A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI

Erico Tjoa and Cuntai Guan, *Fellow, IEEE*

***Abstract*— Recently, artificial intelligence and machine learning in general have demonstrated remarkable performances in many tasks, from image processing to natural language processing, especially with the advent of deep learning (DL). Along with research progress, they have encroached upon many different fields and disciplines. Some of them require high level of accountability and thus transparency, for example, the medical sector. Explanations for machine decisions and predictions are thus needed to justify their reliability. This requires greater interpretability, which often means we need to understand the mechanism underlying the algorithms. Unfortunately, the blackbox nature of the DL is still unresolved, and many machine decisions are still poorly understood. We provide a review on interpretabilities suggested by different research works and categorize them. The different categories show different dimensions in interpretability research, from approaches that provide "obviously" interpretable information to the studies of complex patterns. By applying the same categorization to interpretability in medical research, it is hoped that: 1) clinicians and practitioners can subsequently approach these methods with caution; 2) insight into interpretability will be born with more considerations for medical practices; and 3) initiatives to push forward data-based, mathematically grounded, and technically grounded medical education are encouraged.**

***Index Terms*— Explainable artificial intelligence (XAI), interpretability, machine learning (ML), medical information system, survey.**

I. Introduction

MACHINE learning (ML) has grown large in both research and industrial applications, especially with the success of deep learning (DL) and neural networks (NNs), so large that its impact and possible after-effects can no longer be taken for granted. In some fields, failure is not an option: even a momentarily dysfunctional computer vision algorithm in autonomous vehicle easily leads to fatality. In the medical field, clearly human lives are on the line. Detection of a disease at its early phase is often critical to the recovery of patients or to prevent the disease from advancing to more severe stages. While ML methods, artificial NNs, brain–machine interfaces, and related subfields have recently demonstrated promising performance in performing medical tasks, they are hardly perfect [1]–[9].

Interpretability and explainability of ML algorithms have thus become pressing issues: who is accountable if things go wrong? Can we explain why things go wrong? If things are working well, do we know why and how to leverage them further? Many articles have suggested different measures and frameworks to capture interpretability, and the topic explainable artificial intelligence (XAI) has become a hotspot in ML research community. Popular DL libraries have started to include their own XAI libraries, such as Pytorch Captum and tensorflow tf-explain. Furthermore, the proliferation of interpretability assessment criteria (such as reliability, causality, and usability) helps ML community keep track of how algorithms are used and how their usage can be improved, providing guiding posts for further developments [10]–[12]. In particular, it has been demonstrated that visualization is capable of helping researchers detect erroneous reasoning in classification problems that many previous researchers possibly have missed [13].

The above said, there seems to be a lack of uniform adoption of interpretability assessment criteria across the research community. There have been attempts to define the notions of "interpretability," "explainability" along with "reliability," "trustworthiness," and other similar notions without clear expositions on how they should be incorporated into the great diversity of implementations of ML models; consider [10] and [14]–[18]. In this survey, we will instead use "explainability" and "interpretability" interchangeably, considering a research to be related to interpretability if it does show any attempts: 1) to explain the decisions made by algorithms; 2) to uncover the patterns within the inner mechanism of an algorithm; and 3) to present the system with coherent models or mathematics, and we will include even loose attempts to raise the credibility of machine algorithms.

In this work, we survey through research works related to the interpretability of ML or computer algorithms in general, categorize them, and then apply the same categories to interpretability in the medical field. The categorization is especially aimed to give clinicians and practitioners a perspective on the use of interpretable algorithms that are available in diverse forms. The tradeoff between the ease of interpretation and the need for specialized mathematical knowledge may create a bias in preference for one method when compared to another without justification based on medical practices. This may further provide a ground for specialized education in the medical sector that is aimed to realize the potentials that

Manuscript received October 15, 2019; revised June 7, 2020 and August 10, 2020; accepted September 24, 2020. Date of publication October 20, 2020; date of current version October 28, 2021. This work was supported by the Health-AI Division, DAMO Academy, Alibaba Group Holding Ltd., through the Alibaba-NTU Talent Program. *(Corresponding author: Erico Tjoa.)*

Erico Tjoa was with the HealthTech Division, Alibaba Group Holding Ltd., Hangzhou 311121, China. He is now with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ericotjo001@e.ntu.edu.sg).

Cuntai Guan is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798.

Digital Object Identifier 10.1109/TNNLS.2020.3027314

TABLE I

LIST OF JOURNAL ARTICLES ARRANGED ACCORDING TO THE INTERPRETABILITY METHODS USED, HOW INTERPRETABILITY IS PRESENTED OR THE SUGGESTED MEANS OF INTERPRETABILITY. THE TABULATION PROVIDES A NONEXHAUSTIVE OVERVIEW OF INTERPRETABILITY METHODS, PLACING SOME DERIVATIVE METHODS UNDER THE UMBRELLA OF THE MAIN METHODS THEY DERIVE FROM. HSI: HUMAN STUDY ON INTERPRETABILITY ✓MEANS THERE IS HUMAN STUDY DESIGNED TO VERIFY IF THE SUGGESTED METHODS ARE INTERPRETABLE BY THE HUMAN SUBJECT. ANN: ✓MEANS EXPLICITLY INTRODUCES NEW ARTIFICIAL NN ARCHITECTURE, MODIFIES EXISTING NETWORKS OR PERFORMS TESTS ON NNS

Methods	HSI	ANN	Mechansim		
CAM with global average pooling [42], [91]	✗	✓	Decomposition	Saliency	Perceptive Interpretability
+ Grad-CAM [43] generalizes CAM, utilizing gradient	✓	✓			
+ Guided Grad-CAM and Feature Occlusion [68]	✗	✓			
+ Respond CAM [44]	✗	✓			
+ Multi-layer CAM [92]	✗	✓			
LRP (Layer-wise Relevance Propagation) [13], [53]	✗	✓			
+ Image classifications. PASCAL VOC 2009 etc [45]	✗	✓			
+ Audio classification. AudioMNIST [47]	✗	✓			
+ LRP on DeepLight. fMRI data from Human Connectome Project [48]	✗	✓			
+ LRP on CNN and on BoW(bag of words)/SVM [49]	✗	✓			
+ LRP on compressed domain action recognition algorithm [50]	✗	✗			
+ LRP on video deep learning, *selective relevance method* [52]	✗	✓			
+ BiLRP [51]	✗	✓			
DeepLIFT [57]	✗	✓			
Prediction Difference Analysis [58]	✗	✓			
Slot Activation Vectors [41]	✗	✓			
PRM (Peak Response Mapping) [59]	✗	✓			
LIME (Local Interpretable Model-agnostic Explanations) [14]	✓	✓	Sensitivity		
+ MUSE with LIME [85]	✓	✓			
+ Guidelinebased Additive eXplanation optimizes complexity, similar to LIME [93]	✓	✓			
# Also listed elsewhere: [56], [69], [71], [94]	N.A.	N.A.			
Others. Also listed elsewhere: [95]	N.A.	N.A.	Others		
+ Direct output labels. Training NN via multiple instance learning [65]	✗	✓			
+ Image corruption and testing Region of Interest statistically [66]	✗	✓			
+ Attention map with autofocus convolutional layer [67]	✗	✓			
DeconvNet [72]	✗	✓	Inversion	Signal	
Inverting representation with natural image prior [73]	✗	✓			
Inversion using CNN [74]	✗	✓			
Guided backpropagation [75], [91]	✗	✓			
Activation maximization/optimization [38]	✗	✓	Optimization		
+ Activation maximization on DBN (Deep Belief Network) [76]	✗	✓			
+ Activation maximization, multifaceted feature visualization [77]	✗	✓			
Visualization via regularized optimization [78]	✗	✓			
Semantic dictionary [39]	✗	✓			
Network dissection [36], [37]	✓	✓	Others		
Decision trees	N.A.	N.A.	Verbal		
Propositional logic, rule-based [82]	✗	✗			
Sparse decision list [83]	✗	✗			
Decision sets, rule sets [84], [85]	✓	✗			
Encoder-generator framework [86]	✗	✓			
Filter Attribute Probability Density Function [87]	✗	✗			
MUSE (Model Understanding through Subspace Explanations) [85]	✓	✓			

reside within these algorithms. We also find that many journal articles in the ML and AI community are algorithm-centric. They often assume that the algorithms used are obviously interpretable without conducting human subject tests to verify their interpretability (see column HSI of Tables I and II). It is noted that assuming that a model is obviously interpretable is not necessarily wrong, and, in some cases human tests might be irrelevant (for example, predefined models based on commonly accepted knowledge specific to the content-subject may be considered interpretable without human subject tests). In the tables, we also include a column to indicate whether the interpretability method applies for artificial NN, since the issue of interpretability is recently gathering attention due to its blackbox nature.

We will not attempt to cover all related works many of which are already presented in the research articles and survey we cite [1], [2], [15]–[30]. We extend the so-called integrated interpretability [16] by including considerations for subject-content-dependent models. Compared to [17], we also overview the mathematical formulation of common or popular methods, revealing the great variety of approaches to interpretability. Our categorization draws a starker borderline between the different views of interpretability that seem to be difficult to reconcile. In a sense, our survey is more suitable for technically oriented readers due to some mathematical details, although casual readers may find useful references for relevant popular items, from which they may develop interests in this young research field. Conversely, algorithm users that need interpretability in their work might develop an inclination to understand what is previously hidden in the thick veil of mathematical formulation, which might ironically undermine reliability and interpretability. Clinicians and medical practitioners already having some familiarity with mathematical terms may get a glimpse on how some proposed interpretability

TABLE II
(CONTINUED FROM TABLE I) LIST OF JOURNAL ARTICLES ARRANGED ACCORDING TO THE INTERPRETABILITY METHODS USED, HOW INTERPRETABILITY IS PRESENTED OR THE SUGGESTED MEANS OF INTERPRETABILITY

Methods	HSI	ANN	Mechanism		
Linear probe [101]	✗	✓	Pre-defined models		Interpretability via Mathematical Structure
Regression based on CNN [106]	✗	✓			
Backwards model for interpretability of linear models [107]	✗	✗			
GDM (Generative Discriminative Models): ridge regression + least square [100]	✗	✗			
GAM, GA^{2}M (Generative Additive Model) [82], [102], [103]	✗	✗			
ProtoAttend [105]	✗	✓			
Other content-subject-specific models:	N.A.	N.A.			
+ Kinetic model for CBF (cerebral blood flow) [131]	N.A.	✓			
+ CNN for PK (Pharmacokinetic) modelling [132]	N.A.	✓			
+ CNN for brain midline shift detection [133]	N.A.	✓			
+ Group-driven RL (reinforcement learning) on personalized healthcare [134]	N.A.	✓			
+ Also see [108]–[112]	N.A.	✓			
PCA (Principal Components Analysis), SVD (Singular Value Decomposition)	N.A.	N.A.	Correlation	Feature Extraction	
CCA (Canonical Correlation Analysis) [113]	✗	✗			
SVCCA (Singular Vector Canonical Correlation Analysis) [97] = CCA+SVD	✗	✓			
F-SVD (Frame Singular Value Decomposition) [114] on electromyography data	✗	✗			
DWT (Discrete Wavelet Transform) + Neural Network [135]	✗	✓			
MODWPT (Maximal Overlap Discrete Wavelet Package Transform) [136]	✗	✗			
GAN-based Multi-stage PCA [118]	✓	✗			
Estimating probability density with deep feature embedding [119]	✗	✓			
t-SNE (t-Distributed Stochastic Neighbour Embedding) [77]	✗	✓	Clustering		
+ t-SNE on CNN [120]	✗	✓			
+ t-SNE, activation atlas on GoogleNet [121]	✗	✓			
+ t-SNE on latent space in meta-material design [122]	✗	✓			
+ t-SNE on genetic data [137]	✗	✓			
+ mm-t-SNE on phenotype grouping [138]	✗	✓			
Laplacian Eigenmaps visualization for Deep Generative Model [124]	✗	✓			
KNN (k-nearest neighbour) on multi-center low-rank rep. learning (MCLRR) [125]	✗	✓			
KNN with triplet loss and *query-result activation map pair* [139]	✗	✓			
Group-based Interpretable NN with RW-based Graph Convolutional Layer [123]	✗	✓			
TCAV (Testing with Concept Activation Vectors) [96]	✓	✓	Sensitivity		
+ RCV (Regression Concept Vectors) uses TCAV with Br score [140]	✗	✓			
+ Concept Vectors with UBS [141]	✗	✓			
+ ACE (Automatic Concept-based Explanations) [56] uses TCAV	✓	✓			
Influence function [129] helps understand adversarial training points	✗	✓			
Representer theorem [130]	✗	✓			
SocRat (Structured-output Causual Rationalizer) [127]	✗	✓			
Meta-predictors [126]	✗	✓			
Explanation vector [128]	✗	✗			
# Also listed elsewhere: [14], [43], [85], [94]	N.A.	N.A.			
# Also listed elsewhere: [14], [60], [85] etc	N.A.	N.A.	Optimization		
CNN with separable model [142]	✗	✓	Others		
Information theoretic: Information Bottleneck [98], [99]	✗	✓			
Database of methods v.s. interpretability [10]	N.A.	N.A.	Data Driven		Other Persp.
Case-Based Reasoning [143]	✓	✗			
Integrated Gradients [69], [94]	✗	✓	Invariance		
Input invariance [71]	✗	✓			
Application-based [144], [145] Human-based [146], [147] Function-based [2], [5], [42]–[44], [96], [97], [144], [145]	N.A.	N.A.	Utilities		

methods might be risky and unreliable. The survey [30] views interpretability in terms of extraction of relational knowledge, more specifically, by scrutinizing the methods under neural-symbolic cycle. It presents the framework as a subcategory within the interpretability literature. We include it under verbal interpretability, though the framework does demonstrate that methods in other categories can be perceived under verbal interpretability as well. The extensive survey [18] provides a large list of researches categorized under transparent model and models requiring *post hoc* analysis with multiple subcategories. Our survey, on the other hand, aims to overview the state of interpretable ML as applied to the medical field.

This article is arranged as the following. Section II introduces generic types of interpretability and their subtypes. In each section, where applicable, we provide challenges and future prospects related to the category. Section III applies the categorization of interpretabilities in Section II to medical field and lists a few risks of machine interpretability in the medical field. Before we proceed, it is also imperative to point out that the issue of accountability and interpretability has spawned discussions and recommendations [31]–[33], and even entered the sphere of ethics and law enforcements [34], engendering movements to protect the society from possible misuses and harms in the wake of the increasing use of AI.

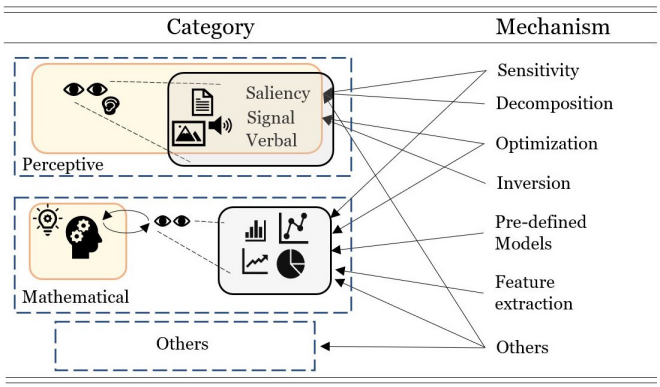

Fig. 1. Overview of categorization with illustration. Orange box: interpretability interface to demarcate the separation between interpretable information and the cognitive process required to understand them. Gray box: algorithm output/product that is proposed to provide interpretability. Black arrow: computing or comprehension process. The perceptive interpretability methods generate items that are usually considered immediately interpretable. On the other hand, methods that provide interpretability via mathematical structure generate outputs that require one more layer of cognitive processing interface before reaching the interpretable interface. The eyes and ear icons represent human senses interacting with items generated for interpretability.

II. Types of Interpretability

There has yet to be a widely adopted standard to understand ML interpretability, though there have been works proposing frameworks for interpretability [10], [13], [35]. In fact, different works use different criteria, and they are justifiable in one way or another. Network dissection has been suggested [36], [37] to evaluate the interpretability of visual representations in deep NN (DNN) inspired by neuroscientists' procedures to understand biological neurons. The articles also offer a way to quantify neuronal network units' activation in response to different concepts detected. The interactive websites [38], [39] have suggested a unified framework to study interpretabilities that have thus-far been studied separately. The article [40] defines a unified measure of feature importance in the SHapley Additive exPlanations (SHAP) framework. Here, we categorize existing interpretabilities and present a nonexhaustive list of works in each category.

The two major categories presented here, namely perceptive interpretability and interpretability by mathematical structures, as illustrated in Fig. 1, appear to present different polarities within the notion of interpretability. An example of the difficulty with perceptive interpretability is as the following. When a visual "evidence" is given erroneously, the algorithm or method used to generate the "evidence" and the underlying mathematical structure sometimes do not offer any useful clues on how to fix the mistakes. On the other hand, a mathematical analysis of patterns may provide information in high dimensions. They can only be easily perceived once the pattern is brought into lower dimensions, abstracting some fine-grained information we could not yet prove is not discriminative with measurable certainty.

A. Perceptive Interpretability

We include in this category interpretabilities that can be humanly perceived, often one that will be considered "obvious." For example, as shown in Fig. 2(a2), an algorithm that classifies an image into the "cat" category can be considered obviously interpretable if it provides segmented patch showing the cat as the explanation. We should note that this alone might on the other hand be considered insufficient, because it: 1) still does not unblackbox an algorithm and 2) ignores the possibility of using background objects for its decision. The following are the subcategories to perceptive interpretability. Refer to Fig. 3 for the overview of the common subcategories.

1) Saliency: Saliency method explains the decision of an algorithm by assigning values that reflect the importance of input components in their contribution to that decision. These values could take the forms of probabilities and super-pixels such as heatmaps etc. For example, Fig. 2(a1) shows how a model predicts that the patient suffers from flu from a series of factors, but LIME [14] explains the choice by highlighting the importance of the particular symptoms that indicate that the illness should indeed be flu. Similarly, Jacovi *et al.* [41] computes the scores reflecting the n-grams activating convolution filters in natural language processing (NLP). Fig. 2(a2) demonstrates the output that LIME will provide as the explanation for the choice of classifications "cat" and Fig. 2(a3) demonstrates a kind of heatmap that shows the contribution of pixels to the segmentation result (segmentation result not shown, and this figure is only for demonstration). More formally, given that model f makes a prediction $y = f(x)$ for input x, for some metric v, typically large magnitude of $v(x_i)$ indicates that the component x_i is a significant reason for the output y.

Saliency methods via decomposition have been developed. In general, they decompose signals propagated within their algorithm and selectively rearrange and process them to provide interpretable information. Class activation map (CAM) has been a popular method to generate heat/saliency/relevance-map (from now, we will use the terms interchangeably) that corresponds to discriminative features for classifications [42]–[44]. The original implementation of CAM [42] produces heatmaps using $f_k(x, y)$, the pixel-wise activation of unit k across spatial coordinates (x, y) in the last convolutional layers, weighted by w_k^c, the coefficient corresponding to unit k for class c. CAM at pixel (x, y) is thus given by $M_c(x, y) = \Sigma_k w_k^c f_k(x, y)$.

Similarly, widely used layer-wise relevance propagation (LRP) is introduced in [45]. Some articles that use LRP to construct saliency maps for interpretability include [13] and [46]–[51]. It is also applicable for video processing [52]. A short summary for LRP is given in [53]. LRP is considered a decomposition method [54]. Indeed, the importance scores are decomposed such that the sum of the scores in each layer will be equal to the output. In short, the relevance score is the pixel-wise intensity at the input layer $R^{(0)}$ where $R_i^{(l)} = \Sigma_j((a_i^{(l)} w_{ij}^+)/(\Sigma_i a_i^{(l)} w_{ij}^+))R_j^{(l+1)}$ is the relevance score of neuron i at layer l with the input layer being at $l = 0$. Each pixel (x, y) at the input layer is assigned the importance value $R^{(0)}(x, y)$, although some combinations of relevance scores $\{R_c^{(l)}\}$ at inner layer l over different channels $\{c\}$ have been demonstrated to be meaningful as well (though possibly less precise; see the tutorial in its website heatmapping.org). LRP can be understood in deep Taylor decomposition

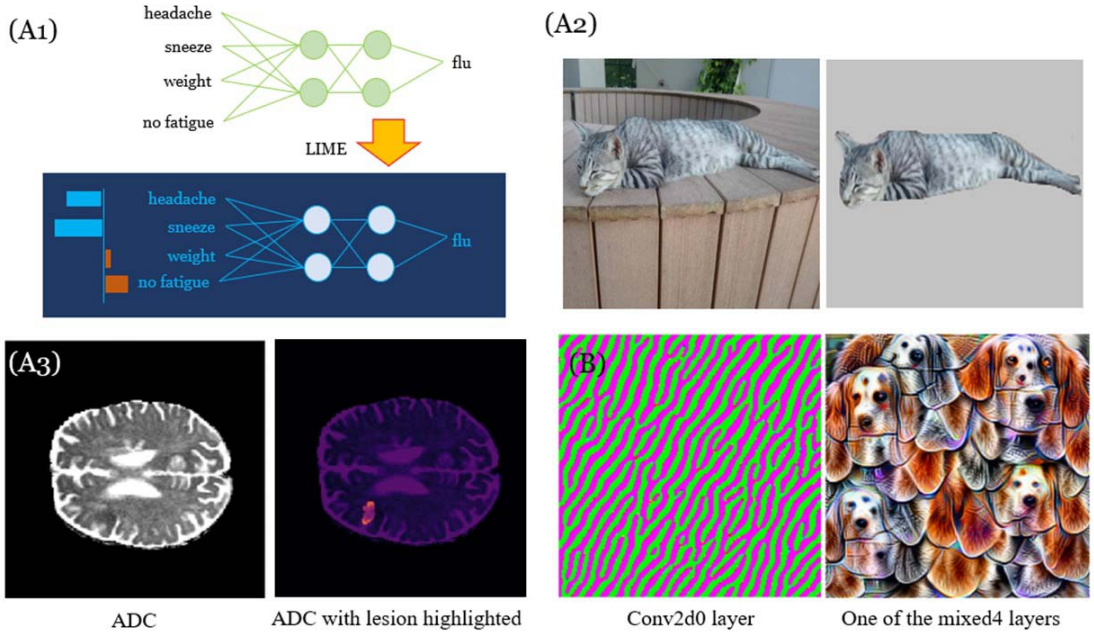

Fig. 2. (a1) Using LIME to generate explanation for text classification. Headache and sneeze are assigned positive values. This means both factors have positive contribution to the model prediction "flu." On the other hand, weight and no fatigue contribute negatively to the prediction. (a2) LIME is used to generate the super-pixels for the classification "cat." (a3) ADC modality of a slice of MRI scan from ISLES 2017 segmentation competition. Reddish intensity region reflects a possible "explanation" to the choice of segmentation (segmentation not shown). (b) Optimized images that maximize the activation of a neuron in the indicated layers. In shallower layer, simple patterns activate neurons strongly while in deeper layer, more complex features such as dog faces and ears do. Figure (b) is obtained from https://distill.pub/2018/building-blocks/ with permission from Chris Olah.

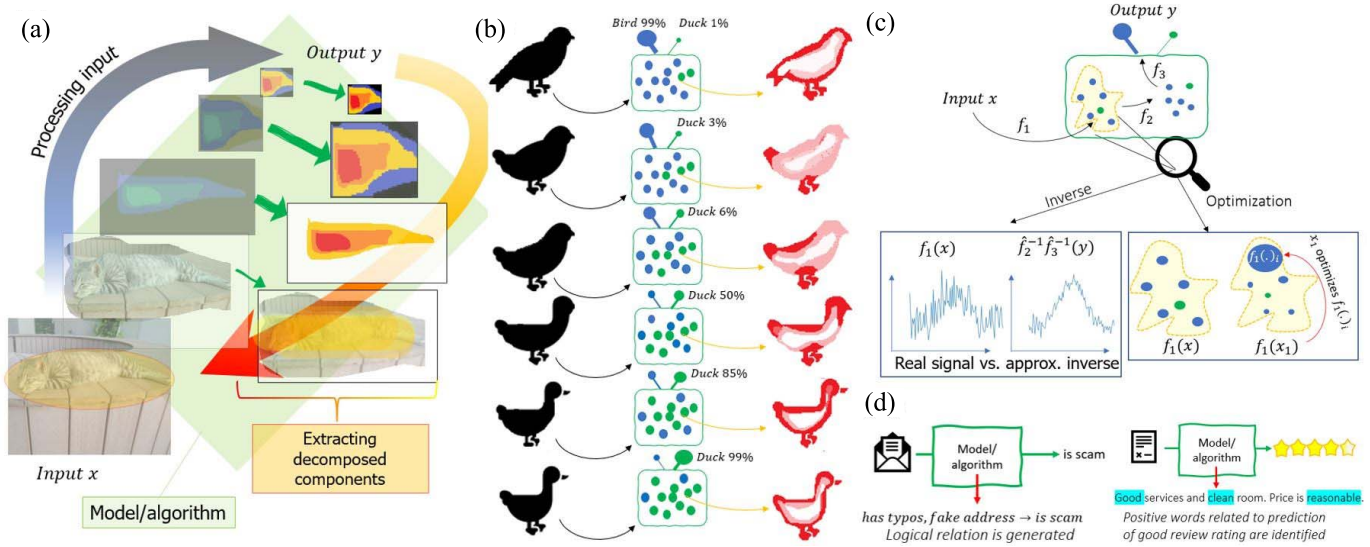

Fig. 3. Overview on perceptive interpretability methods. (a) Saliency method with decomposition mechanism. The input which is an image of a cat is fed into the model for processing along the blue arrow. The resulting output and intermediate signals (green arrows) are decomposed and selectively picked for processing, hence providing information for the intermediate mechanism of the model in the form of (often) heatmappings, shown in red/orange/yellow colors. (b) Saliency method with sensitivity mechanism. The idea is to show how small changes to the input (black figures of birds and ducks) affect the information extracted for explainability (red silhouette). In this example, red regions indicate high relevance, which we sometimes observe at edges or boundary of objects, where gradients are high. (c) Signal method by inversion and optimization. Inverses of signals or data propagated in a model could possibly reveal more sensible information (see arrow labeled "inverse"). Adjusting input to optimize a particular signal (shown as the ith component of the function f_1) may provide us with x_1 that reveals explainable information (see arrow labeled "optimization"). For illustration, we show that the probability of correctly predicting duck improves greatly once the head is changed to the head of a duck which the model recognizes. (d) Verbal interpretability is typically achieved by ensuring that the model is capable of providing humanly understable statements, such as the logical relation or the positive words shown.

framework [55], though, as we speak, many versions of LRP are being developed. The code implementation can also be found in the aforementioned website.

Automatic concept-based explanations (ACEs) algorithm [56] uses super-pixels as explanations. Other decomposition methods that have been developed include, DeepLIFT and gradient*input [57], prediction difference analysis [58] and [41]. Peak response mapping [59] is generated by backpropagating peak signals. Peak signals are normalized and treated as probability, and the method can be seen as decomposition

into probability transitions. In [60], removed correlation ρ is proposed as a metric to measure the quality of signal estimators. And then it proposes PatternNet and PatternAttribution that backpropagate parameters optimized against ρ, resulting in saliency maps as well. SmoothGrad [61] improves gradient-based techniques by adding noises. Do visit the related website that displays numerous visual comparison of saliency methods; be mindful of how some heatmaps highlight apparently irrelevant regions.

For NLP or sentiment analysis, saliency map can also take the form of "heat" scores over words in texts, as demonstrated by Arras *et al.* [62] using LRP and by Karpathy *et al.* [63]. In the medical field (see later section), Irvin *et al.* [6], Zhao *et al.* [44], Paschali *et al.* [64], Couture *et al.* [65], Li *et al.* [66], Qin *et al.* [67], Tang *et al.* [68], Papanastasopoulos *et al.* [69], and Lee *et al.* [70] have studied methods employing saliency and visual explanations. It is noted that we also subcategorize LIME as a method that uses optimization and sensitivity as its underlying mechanisms, and many researches on interpretability span more than one subcategories.

a) Challenges and future prospects: As seen, the formulas for CAM and LRP are given on a heuristic: certain ways of interaction between weights and the strength of activation of some units within the models will eventually produce the interpretable information. The intermediate processes are not amenable to scrutiny. For example, taking one of the weights and changing its value does not easily reveal any useful information. How these prescribed ways translate into interpretable information may also benefit from stronger evidences, especially evidences beyond visual verification of localized objects. Signal methods to investigate ML models (see later section) exist, but such methods that probe them with respect to the above methods have not been attempted systematically, possibly opening up a different research direction.

2) Signal Method: Methods of interpretability that observe the stimulation of neurons or a collection of neurons are called signal methods [71]. On the one hand, the activated values of neurons can be manipulated or transformed into interpretable forms. For example, the activation of neurons in a layer can be used to reconstruct an image similar to the input. This is possible because neurons store information systematically [36], [72]: feature maps in the deeper layer activate more strongly to complex features, such as human face, keyboard, etc., while feature maps in the shallower layers show simple patterns such as lines and curves. An example of feature map is the output of a convolutional filter in a convolutional NN (CNN). Network dissection procedure evaluates neuronal unit's activation by computing its IoU score that is relevant to a concept in question [36], [37]. On the other hand, parameters or even the input data might be optimized with respect to the activation values of particular neurons using methods known as activation optimization (see a later section). The following are the relevant subcategories.

a) Feature maps and inversions for input reconstructions: A feature map often looks like a highly blurred image with most region showing zero (or low intensity), except for the patch that a human could roughly discern as a detected feature. Sometimes, these discernible features are considered interpretable, as in [72]. However, they might be too distorted.

Then, how else can a feature map be related to a humanly perceptible feature? An inverse convolution map can be defined: for example, if feature map in layer 2 is computed in the network via $y_2 = f_2(f_1(x))$ where x is the input, $f_1(.)$ consists of 7×7 convolutions of stride 2 followed by max-pooling and likewise $f_2(.)$. Then [72] reconstructs an image using a deconvolution network by approximately inversing the trained convolutional network $\tilde{x} = \text{deconv}(y) = \hat{f}_2^{-1}\hat{f}_1^{-1}(y)$ which is an approximation, because layers such as max-pooling have no unique inverse. It is shown that $\tilde{x}$ does appear like slightly blurred version of the original image, which is distinct to human eye. Inversion of image representations within the layers has also been used to demonstrate that CNN layers do store important information of an input image accurately [73], [74]. Guided backpropagation [75] modifies the way backpropagation is performed to achieve inversion by zeroing negative signals from both the output or input signals backwards through a layer. Indeed, inversion-based methods do use saliency maps for visualization of the "activated" signals.

b) Activation optimization: Besides transforming the activation of neurons, signal method also includes finding input images that optimize the activation of a neuron or a collection of neurons. This is called the activation maximization. Starting with a noise as an input x, the noise is slowly adjusted to increase the activation of a select (collection of) neuron(s) $\{a_k\}$. In simple mathematical terms, the task is to find $x_0 = \text{argmax } ||\{a_k\}||$ where optimization is performed over input x and $||.||$ is a suitable metric to measure the combined strength of activations. Finally, the optimized input that maximizes the activation of the neuron(s) can emerge as something visually recognizable. For example, the image could be a surreal fuzzy combination of swirling patterns and parts of dog faces, as shown in Fig. 2(b).

Research works on activation maximization include [76] on MNIST data set, [77] and [78] that uses a regularization function. In particular, Olah *et al.* [38] provides an excellent interactive interface (feature visualization) demonstrating activation-maximized images for GoogLeNet [79]. GoogLeNet has a deep architecture, from which we can see how neurons in deeper layer stores complex features while shallower layer stores simple patterns [see Fig. 2(b)]. To bring this one step further, the "semantic dictionary" is used [39] to provide a visualization of activations within a higher level organization and semantically more meaningful arrangements.

c) Other observations of signal activations: Ablation studies [80], [81] also study the roles of neurons in shallower and deeper layers. In essence, some neurons are corrupted and the output of the corrupted NN is compared to the original network.

d) Challenges and future prospects: Signal methods might have revealed some parts of the black-box mechanisms. Many questions still remain which are as follows.

1) What do we do with the (partially) reconstructed images and images that optimize activation?

2) We might have learned how to approximately inverse signals to recover images, can this help improve interpretability further?
3) The components and parts in the intermediate process that reconstruct the approximate images might contain important information; will we be able to utilize them in the future?
4) How is explaining the components in this "inverse space" more useful than explaining signals that are forward propagated?
5) Similarly, how does looking at intermediate signals that lead to activation optimization help us pinpoint the role of a collection of neurons?
6) Optimization of highly parameterized functions notoriously gives nonunique solutions. Can we be sure that optimization that yields combination of surreal dog faces will not yield other strange images with minor alteration?

In the process of answering these questions, we may find hidden clues required to get closer to interpretable AI.

3) Verbal Interpretability: This form of interpretability takes the form of verbal chunks that human can grasp naturally. Examples include sentences that indicate causality, as shown in the examples below.

Logical statements can be formed from proper concatenation of predicates, connectives, etc. An example of logical statement is the conditional statement. Conditional statements are statements of the form $A \rightarrow B$, in another words "if A then B." An ML model from which logical statements can be extracted directly has been considered obviously interpretable. The survey [30] shows how interpretability methods in general can be viewed under such symbolic and relational system. In the medical field, see [82], [83].

Similarly, decision sets or rule sets have been studied for interpretability [84]. The following is a single line in a rule set "rainy and grumpy or calm $\rightarrow$ dairy or vegetables," directly quoted from the article. Each line in a rule set contains a clause with an input in disjunctive normal form (DNF) mapped to an output in DNF as well. The example above is formally written (rainy $\wedge$ grumpy) $\vee$ calm $\rightarrow$ dairy $\vee$ vegetables. Comparing three different variables, it is suggested that interpretability of explanations in the form of rule sets is most affected by cognitive chunks, explanation size and little effected by variable repetition. Here, a cognitive chunk is defined as a clause of inputs in DNF and the number of (repeated) cognitive chunks in a rule set is varied. The explanation size is self-explanatory (a longer/shorter line in a rule set, or more/less lines in a rule set). MUSE [85] also produces explanation in the form of decision sets, where interpretable model is chosen to approximate the black-box function and optimized against a number of metrics, including direct optimization of interpretability metrics.

It is not surprising that verbal segments are provided as the explanation in NLP problems. An encoder-generator framework [86] extracts segment like "a very pleasant ruby red-amber color" to justify 5 out of 5-star rating for a product review. Given a sequence of words $x = (x_1, \ldots, x_l)$ with $x_k \in \mathbb{R}^d$, explanation is given as the subset of the sentence that gives a summary of why the rating is justified. The subset can be expressed as the binary sequence $(z_1, \ldots, z_l)$ where $z_k = 1(0)$ indicates x_k is (not) in the subset. Then z follows a probability distribution with $p(z|x)$ decomposed by assuming independence to $\Pi_k p(z_k|x)$ where $p(z_k|x) = \sigma_z(W^z[\overrightarrow{h_k}, \overleftarrow{h_k}] + b^z)$, with $\overrightarrow{h_t}$, $\overleftarrow{h_t}$ being the usual hidden units in the recurrent cell (forward and backward, respectively). Similar segments are generated using filter-attribute probability density function to improve the relation between the activation of certain filters and specific attributes [87]. Earlier works on visual question answering (VQA) [88]–[90] are concerned with the generation of texts discussing objects appearing in images.

a) Challenges and future prospects: While texts appear to provide explanations, the underlying mechanisms used to generate the texts are not necessarily explained. For example, NNs and the common variants/components used in text-related tasks such as recurrent NN (RNN), long short-term memory (LSTM) are still black boxes that are hard to troubleshoot in the case of wrong predictions. There have been less works that probe into the inner signals of LSTM and RNN NNs. This is a possible research direction, although similar problem as mentioned in Section II-A2d may arise (what to do with the intermediate signals?). Furthermore, while word embedding is often optimized with the usual loss minimization, there does not seem to be a coherent explanation to the process and shape of the optimized embedding. There may be some clues regarding optimization residing within the embedding, and thus successfully interpreting the shape of embedding may help shed light into the mechanism of the algorithm.

B. Interpretability via Mathematical Structure

Mathematical structures have been used to reveal the mechanisms of ML and NN algorithms. In the previous section, deeper layer of NN is shown to store complex information while shallower layer stores simpler information [72]. Testing with concept activation vector (TCAV) [96] has been used to show similar trend, as suggested in Fig. 4(a2). Other methods include clustering, such as t-distributed stochastic neighbor embedding (t-SNE) shown in Fig. 4(b) and subspace-related methods, for example correlation-based singular vector canonical correlation analysis (SVCCA) [97] is used to find the significant directions in the subspace of input for accurate prediction, as shown in Fig. 4(c). Information theory has been used to study interpretability by considering Information Bottleneck principle [98], [99]. The rich ways in which mathematical structures add to the interpretability pave ways to a comprehensive view of the interpretability of algorithms, hopefully providing a ground for unifying the different views under a coherent framework in the future. Fig. 5 provides an overview of ideas under this category.

1) Predefined Model: To study a system of interest, especially complex systems with not well-understood behavior, mathematical formula such as parametric models can help simplify the tasks. With a proper hypothesis, relevant terms and parameters can be designed into the model. Interpretation of the terms come naturally if the hypothesis is either consistent with available knowledge or at least developed

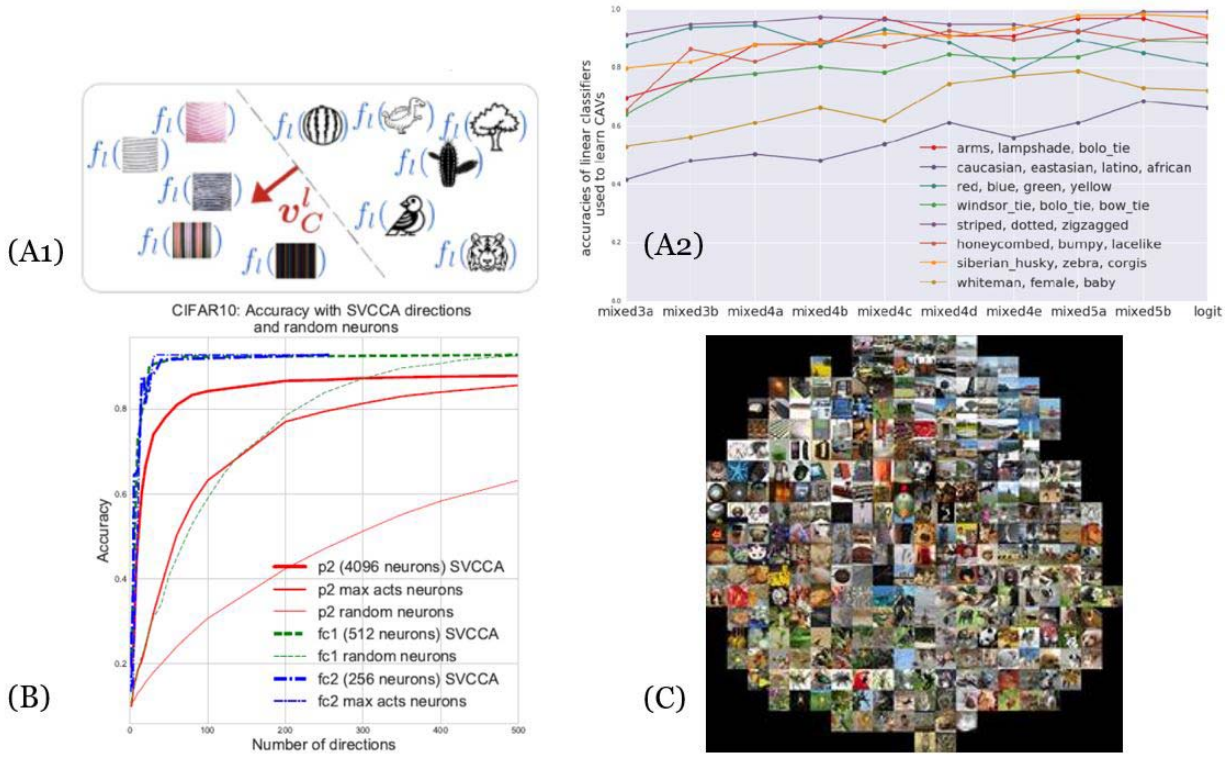

Fig. 4. (a1) TCAV [96] method finds the hyperplane CAV that separates concepts of interest. (a2) Accuracies of CAV applied to different layers supports the idea that deeper NN layers contain more complex concepts, and shallower layers contain simpler concepts. (b) SVCCA [97] finds the most significant subspace (direction) that contains the most information. The graph shows that as few as 25 directions out of 500 are enough to produce the accuracies of the full network. (c) t-SNE clusters images in meaningful arrangement, for example, dog images are close together. Figures (a1) and (a2) are used with permission from the authors Been Kim; figure (b) and (c) from Maithra Raghu and Jascha Sohl-dickstein.

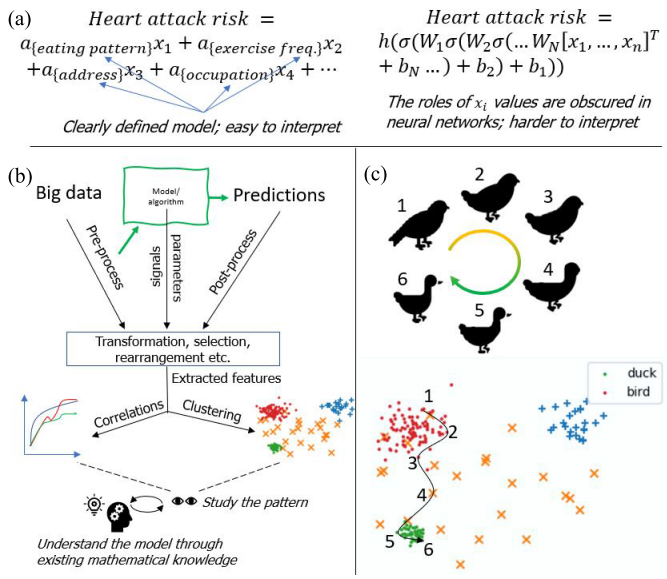

Fig. 5. Overview of methods whose interpretability depend on interpreting underlying mathematical structure. (a) Predefined models. Modeling with clear, easily understandable model, such as linear model can help improve readability, and hence interpretability. On the other hand, using NN could obscure the meaning of input variables. (b) Feature extraction. Data, predicted values, signals, and parameters from a model are processed, transformed, and selectively picked to provide useful information. Mathematical knowledge is usually required to understand the resulting pattern. (c) Sensitivity. Models that rely on sensitivity, gradients, perturbations, and related concepts will try to account for how different data are differently represented. In the figure, the small changes transforming the bird to the duck can be traced along a map obtained using clustering.

with good reasons. When the systems are better understood, these formula can be improved by the inclusion of more complex components. In the medical field (see later section), an example is kinetic modeling. ML can be used to compute the parameters defined in the models. Other methods exist, such as integrating commonly available methodologies with subject specific contents, etc. For example, generative discriminative models [100], combine ridge regression and least square method to handle variables for analyzing Alzheimer's disease and schizophrenia.

a) Linearity: The simplest interpretable predefined model is the linear combination of variables $y = \Sigma_i a_i x_i$, where a_i is the degree of how much x_i contributes to the prediction y. A linear combination model with $x_i \in \{0, 1\}$ has been referred to as the additive feature attribution method [40]. If the model performs well, this can be considered highly interpretable. However, many models are highly nonlinear. In such cases, studying interpretability via linear properties (for example, using linear probe; see below) are useful in several ways, including the ease of implementation. When linear property appears to be insufficient, nonlinearity can be introduced; it is typically not difficult to replace the linear component $\vec{w} \cdot \vec{a}$ within the system with a nonlinear version $f(\vec{w}, \vec{a})$.

A linear probe is used in [101] to extract information from each layer in a NN. More technically, assume we have DL classifier $F(x) \in [0, 1]^D$ where $F_i(x) \in [0, 1]$ is the probability that input x is classified into class i out of D classes. Given a set of features H_k at layer k of a NN, then

the linear probe f_k at layer k is defined as a linear classifier $f_k : H_k \rightarrow [0,1]^D$ that is, $f(h_k) = \text{softmax}(Wh_k + b)$. In another words, the probe tells us how well the information from only layer k can predict the output, and each of this predictive probe is a linear classifier by design. The article then shows plots of the error rate of the prediction made by each f_k against k and demonstrates that these linear classifiers generally perform better at deeper layer, that is, at larger k.

b) General additive models: Linear model is generalized by the generalized additive model (GAM) [102], [103] with standard form $g(E[y]) = \beta_0 + \Sigma f_j(x_j)$ where g is the link function. The equation is general, and specific implementations of f_j and link function depend on the task. The familiar general linear model (GLM) is GAM with the specific implementation of linear f_j and g is the identity. Modifications can be duly implemented. As a natural extension to the model, interaction terms between variables $f_{ij}(x_i, x_j)$ are used [104]; we can certainly extend this indefinitely. ProtoAttend [105] uses probabilities as weights in the linear component of the NN. Such model is considered inherently interpretable by the authors. In the medical field, see [82], [100], [106], [107].

c) Content-subject-specific model: Some algorithms are considered obviously interpretable within its field. Models are designed based on existing knowledge or empirical evidence, and thus interpretation of the models is innately embedded into the system. ML algorithms can then be incorporated in rich and diverse ways, for example, through parameter fitting. The following lists just a few works to illustrate the usage diversity of ML algorithms. Deep Tensor NN is used for quantum many-body systems [108]. Atomistic NN architecture for quantum chemistry is used in [109], where each atom is like a node in a graph with a set of feature vectors. The specifics depend on the NN used, but this model is considered inherently interpretable. NN has been used for programmable wireless environments (PWEs) [110]. TS approximation [111] is a fuzzy network approximation of other NNs. The approximate fuzzy system is constructed with choices of components that can be adapted to the context of interpretation. The article itself uses sigmoid-based membership function, which it considers interpretable. A so-called model-based reinforcement learning (RL) is suggested to be interpretable after the addition of high-level knowledge about the system that is realized as Bayesian structure [112].

d) Challenges and future prospects: The challenge of formulating the "correct" model exists regardless of ML trend. It might be interesting if a system is found that is fundamentally operating on a specific ML model. Backpropagation-based DNN itself is inspired by the brain, but they are not operating at fundamental level of similarity (nor is there any guarantee that such model exists). When interpretability is concerned, having fundamental similarity to real, existing systems may push forward our understanding of ML model in unprecedented ways. Otherwise, in the standard uses of ML algorithm, different optimization paradigms are still being discovered. Having optimization paradigm that is specialized for specific models may be contribute to a new aspect of interpretable ML.

2) Feature Extraction: We give an intuitive explanation via a hypothetical example of a classifier for heart-attack prediction. Given, say, 100-D features including eating pattern, job, and residential area of a subject. A kernel function can be used to find out that the strong predictor for heart attack is a 100-D vector which is significant in the following axes: eating pattern, exercise frequency, and sleeping pattern. Then, this model is considered interpretable because we can link heart-attack risk with healthy habits rather than, say socio-geographical factors. More information can be drawn from the next most significant predictor and so on.

a) Correlation: The methods discussed in this section include the use of correlation in a general sense. This will naturally include covariance matrix and correlation coefficients after transformation by kernel functions. A kernel function transforms high-dimensional vectors such that the transformed vectors better distinguish different features in the data. For example, the principal component (PC) analysis transforms vectors into the PCs that can be ordered by the eigenvalues of singular-value-decomposed (SVD) covariance matrix. The PC with the highest eigenvalue is roughly the most informative feature. Many kernel functions have been introduced, including the canonical correlation analysis (CCA) [113]. CCA provides the set of features that transforms the original variables to the pairs of canonical variables, where each pair is a pair of variables that are "best correlated" but not correlated with other pairs. Quoted from [114], "such features can inherently characterize the object and thus it can better explore the insights and finer details of the problems at hand." In the previous sections, interpretability research using correlation includes [60].

SVCCA combines CCA and SVD to analyze interpretability [97]. Given an input data set $X = \{x_1, \ldots, x_m\}$ where each input x_i is possibly multidimensional. Denote the activation of neuron i at layer l as $z_i^l = (z_i^l(x_1), \ldots, z_i^l(x_m))$. It is noted that one such output is defined for the entire input data set. SVCCA finds out the relation between two layers of a network $l_k = \{z_i^{l_k} | i = 1, \ldots, m_k\}$ for $k = 1, 2$ by taking l_1 and l_2 as the input (generally, l_k does not have to be the entire layer). SVCCA uses SVD to extract the most informative components l_k' and uses CCA to transform l_1' and l_2' such that $\bar{l}_1' = W_X l_1'$ and $\bar{l}_2' = W_X l_2'$ have the maximum correlation $\rho = \{\rho_1, \ldots, \rho_{\min(m_1, m_2)}\}$. One of the SVCCA experiments on CIFAR-10 demonstrates that only 25 most-significant axes in l_k' are needed to obtain nearly the full accuracy of a full-network with 512 dimensions. Besides, the similarity between two compared layers is defined to be $\bar{\rho} = (1/(\min(m_1, m_2)))\Sigma_i \rho_i$.

The successful development of generative adversarial networks (GANs) [115]–[117] for generative tasks have spawned many derivative works. GAN-based models have been able to generate new images not distinguishable from synthetic images and perform many other tasks, including transferring style from one set of images to another or even producing new designs for products and arts. Studies related to interpretabilities exist. For example, [118] uses encoder–decoder system to perform multistage PCA. Generative model is used to show that natural image distribution modeled using probability

density is fundamentally difficult to interpret [119]. This is demonstrated through the use of GAN for the estimation of image distribution density. The resulting density shows preferential accumulation of density of images with certain features (for examples, images featuring small object with few foreground distractions) in the pixel space. The article then suggests that interpretability is improved once it is embedded in the deep feature space, for example, from GAN. In this sense, the interpretability is offered by better correlation between the densities of images with the correct identification of the objects. Consider also the GAN-based works they cite.

b) Clustering: Algorithm such as t-SNE has been used to cluster input images based on their activation of neurons in a network [77], [120]. The core idea relies on the distance between objects being considered. If the distance between two objects are short in some measurement space, then they are similar. This possibly appeals to the notion of human learning by the Law of Association. It differs from correlation-based method which provides some metrics that relate the change of one variable with another, where the two related objects can originate from completely different domains; clustering simply presents their similarity, more sensibly in similar domain or in the subsets thereof. In [120], the activations $\{f_{\text{fc7}}(x)\}$ of 4096-D layer fc7 in the CNN are collected over all input $\{x\}$. Then $\{f_{\text{fc7}}(x)\}$ is fed into t-SNE to be arranged and embedded into two dimensions for visualization (each point then is visually represented by the input image x). Activation atlases are introduced in [121], which similarly uses t-SNE to arrange some activations $\{f_{\text{act}}(x)\}$, except that each point is represented by the average activations of feature visualization. In metamaterial design [122], design pattern and optical responses are encoded into latent variables to be characterized by variational auto encoder (VAE). Then, t-SNE is used to visualize the latent space.

In the medical field (also see later section), we have [123], [124] (uses Laplacian eigenmap (LE) for interpretability), and [125] (introduces a low-rank representation method for autistic spectrum diagnosis).

c) Challenges and future prospects: This section exemplifies the difficulty in integrating mathematics and human intuition. Having extracted "relevant" or "significant" features, sometimes we are left with still a combination of high-dimensional vectors. Further analysis comes in the form of correlations or other metrics that attempt to show similarities or proximity. The interpretation may stay as mathematical artifact, but there is a potential that separation of concepts attained by these methods can be used to reorganize a black-box model from within. It might be an interesting research direction that lacks justification in terms of real-life application: however, progress in unraveling black-boxes may be a high-risk high-return investment.

3) Sensitivity: We group together methods that rely on localization, gradients, and perturbations under the category of "sensitivity." These methods rely on the notion of small changes dx in calculus and the neighborhood of a point in metric spaces.

a) Sensitivity to input noises or neighborhood of data points: Some methods rely on the locality of some input x. Let a model $f(.)$ predicts $f(x)$ accurately for some x. Denote $x+\delta$ as a slightly noisy version of x. The model is locally faithful if $f(x+\delta)$ produces correct prediction, otherwise, the model is unfaithful and clearly such instability reduces its reliability. Fong and Vedaldi [126] introduces meta-predictors as interpretability methods and emphasizes the importance of the variation of input x to NN in explaining a network. They define explanation and local explanation in terms of the response of blackbox f to some input. Amongst many of the studies conducted, they provide experimental results on the effect of varying input such as via deletion of some regions in the input. Likewise, when random pixels of an image are deleted (hence the data point is shifted to its neighborhood in the feature space) and the resulting change in the output is tested [57], pixels that are important to the prediction can be determined. In text classification, Alvarez-Melis and Jaakkola [127] provides "explanations" in the form of partitioned graphs. The explanation is produced in three main steps, where the first step involves sampling perturbed versions of the data using VAE.

TCAVs has also been introduced as a technique to interpret the low-level representation of NN layer [96]. First, the concept activation vector (CAV) is defined. Given input $x \in \mathbb{R}^n$ and a feedforward layer l having m neurons, the activation at that layer is given by $f_l : \mathbb{R}^n \to \mathbb{R}^m$. If we are interested in the concept C, for example "striped" pattern, then, using TCAV, we supply a set P_C of examples corresponding to "striped" pattern (zebra, clothing pattern, etc.) and the negative examples N. This collection is used to train a binary classifier $v_C^l \in \mathbb{R}^m$ for layer l that partitions $\{f_l(x) : x \in P_C\}$ and $\{f_l(x) : x \in N\}$. In another words, a kernel function extracts features by mapping out a set of activations that has relevant information about the "stripe"-ness. CAV is thus defined as the normal vector to the hyperplane that separates the positive examples from the negative ones, as shown in Fig. 4(a1). It then computes directional derivative $S_{v,k,l}(x) = \nabla h_{l,k}(f_l(x)) \cdot v_C^l$ to obtain the sensitivity of the model with respect to the concept C, where $h_{l,k}$ is the logit function for class k of C for layer l.

LIME [14] optimizes over models $g \in G$ where G is a set of interpretable models G by minimizing locality-aware loss and complexity. In another words, it seeks to obtain the optimal model $\xi(x) = \text{argmin}_{g \in G}\ L(f, g, \pi_x) + \Omega(g)$ where Ω is the complexity and f is the true function we want to model. An example of the loss function is $L(f, g, \pi_x) = \Sigma_{z,z' \in Z} \pi_x(z)[f(x) - g(z')]^2$ with $\pi_x(z)$ being, for example, Euclidean distance and Z is the vicinity of x. From the equation, it can be seen that the desired g will be close to f in the vicinity Z of x, because $f(z) \approx g(z')$ for $z, z' \in Z$. In another words, noisy inputs z, z' do not add too much losses.

Gradient-based explanation vector $\xi(x_0) = (\partial/\partial x)P(Y \neq g(x_0)|X = x)$ is introduced in [128] for Bayesian classifier $g(x) = \text{argmin}_{c \in \{i,\ldots,C\}}\ P(Y \neq c|X = x)$, where x, ξ are d-dimensional. For any $i = 1, \ldots, d$, high absolute value of $[\xi(x_0)]_i$ means that component i contributes significantly to the decision of the classifier. If it is positive, the higher the value is, the less likely x_0 contributes to decision $g(x_0)$.

ACE algorithm [56] uses TCAV to compute saliency score and generate super-pixels as explanations. Grad-CAM [43] is

a saliency method that uses gradient for its sensitivity measure. In [129], influence function is used. While theoretical, the article also practically demonstrates how understanding the underlying mathematics will help develop perturbative training point for adversarial attack.

b) Sensitivity to data set: A model is possibly sensitive to the training data set $\{x_i\}$ as well. Influence function is also used to understand the effect of removing x_i for some i and shows the consequent possibility of adversarial attack [129]. Studies on adversarial training examples can be found in the article and its citations, where seemingly random, insignificant noises can degrade machine decision considerably. The representer theorem is introduced for studying the extent of effect x_i has on a decision made by a DNN [130].

c) Challenges and future prospects: There seems to be a concern with locality and globality of the concepts. As mentioned in [96], to achieve global quantification for interpretability, explanation must be given for a set of examples or the entire class rather than "just explain individual data inputs." As a specific example, there may be a concern with the globality of TCAV. From our understanding, TCAV is a perturbation method by the virtue of stable continuity in the usual derivative and it is global because the whole subset of data set with label k of concept C has been shown to be well distinguished by TCAV. However, we may want to point out that despite their claim to globality, it is possible to view the success of TCAV as local, since it is only "global" within each label k rather than within all data set considered at once.

From the point of view of image processing, the neighborhood of a data point (an image) in the feature space poses a rather subtle question; also refer to Fig. 5(c) for related illustration. For example, after rotating and stretching the image or deleting some pixels, how does the position of the image in the feature space change? Is there any way to control the effect of random noises and improve robustness of machine prediction in a way that is sensible to human's perception? The transition in the feature space from one point to another point that belongs to different classes is also unexplored.

On a related note, gradients have played important roles in formulating interpretability methods, be it in image processing or other fields. Current trend recognizes that regions in the input space with significant gradients provide interpretability. Deforming these regions quickly degrades the prediction; conversely, the particular values at these regions are important to the reach a certain prediction. This is helpful, since calculus exists to help analyse gradients. However, this has shown to be disruptive as well. For example, imperceptible noises can degrade prediction drastically (see manipulation of explanations under Section III-D). Since gradient is also in the core of loss optimization, it is a natural target for further studies.

4) Optimization: We have described several researches that seek to attain interpretability via optimization methods. Some have optimization at the core of their algorithm, but the interpretability is left to visual observation, while others optimize interpretability mathematically.

a) Quantitatively maximizing interpretability: To approximate a function f, as previously mentioned, LIME [14] performs optimization by finding optimal model $\xi \in G$ so that $f(z) \approx \xi(z')$ for $z, z' \in Z$ where Z is the vicinity of x, so that local fidelity is said to be achieved. Concurrently, the complexity $\Omega(\xi)$ is minimized. Minimized Ω means the model's interpretability is maximized. MUSE [85] takes in blackbox model, prediction and user-input features to output decision sets based on optimization with respect to fidelity, interpretability, and unambiguity. The available measures of interpretability that can be optimized include size, featureoverlap, etc. (refer to Table II of its Appendix).

b) Activation optimization: Activation optimizations are used in research works such as [38] and [76]–[78] as explained in the previous section. The interpretability relies on direct observation of the neuron-activation-optimized images. While the quality of the optimized images are not evaluated, the fact that parts of coherent images emerge with respect to a (collection of) neuron(s) does demonstrate some organization of information in the NNs.

C. Other Perspectives to Interpretability

There are many other concepts that can be related to interpretability. Selvaraju *et al.* [43] conducted experiments to test the improvements of human performance on a task after being given explanations (in the form of visualization) produced by ML algorithms. We believe this might be an exemplary form of interpretability evaluation. For example, we want to compare ML algorithms $\mathrm{ML_A}$ with $\mathrm{ML_B}$. Say, human subjects are given difficult classification tasks and attain a baseline 40% accuracy. Repeat the task with different set of human subjects, but they are given explanations churned out by $\mathrm{ML_A}$ and $\mathrm{ML_B}$. If the accuracies attained are now 50% and 80%, respectively, then $\mathrm{ML_B}$ is more interpretable.

Even then, if human subjects cannot really explain why they can perform better with the given explanations, then the interpretability may be questionable. This brings us to the question of what kind of interpretability is necessary in different tasks and certainly points to the possibility that there is no need for a unified version of interpretability.

1) Data-Driven Interpretability:

a) Data in catalog: A large amount of data has been crucial to the functioning of many ML algorithms, mainly as the input data. In this section, we mention works that put a different emphasize on the treatment of these data arranged in catalog. In essence, Doshi-Velez and Kim [10] suggests that we create a matrix whose rows are different real-world tasks (e.g., pneumonia detection), columns are different methods (e.g., decision tree with different depths) and the entries are the performance of the methods on some end-task. How can we gather a large collection of entries into such a large matrix? Apart from competitions and challenges, crowd-sourcing efforts will aid the formation of such database [148], [149]. A clear problem is how multidimensional and gigantic such tabulation will become, not to mention that the collection of entries is very likely uncountably many. Formalizing interpretability here means we pick latent dimensions (common criteria) that human can evaluate for example, time constraint or time-spent, cognitive chunks (defined as the basic unit of explanation, also see the definition in [84]), etc. These

dimensions are to be refined along iterative processes as more user inputs enter the repository.

b) Incompleteness: In [10], the problem of incompleteness of problem formulation is first posed as the issue in interpretability. Incompleteness is present in many forms, from the impracticality to produce all test cases to the difficulty in justifying why a choice of proxy is the best for some scenarios. At the end, it suggests that interpretability criteria are to be born out of collective agreements of the majority, through a cyclical process of discoveries, justifications, and rebuttals. In our opinion, a disadvantage is that there is a possibility that no unique convergence will be born, and the situation may aggravate if, say, two different conflicting factions are born, each with enough advocate. The advantage lies in the existence of strong roots for the advocacy of certain choice of interpretability. This prevents malicious intent from tweaking interpretability criteria to suit *ad hoc* purposes.

2) Invariances:

a) Implementation invariance: Sundararajan *et al.* [94] suggests implementation invariance as an axiomatic requirement to interpretability. In the article, it is stated as the following. Define two functionally equivalent functions as f_1, f_2 so that $f_1(x) = f_x(x)$ for any x regardless of their implementation details. Given any two such networks using attribution method, then the attribution functional A will map the importance of each component of an input to f_1 the same way it does to f_2. In another words, $(A[f_1](x))_j = (A[f_2](x))_j$ for any $j = 1, \ldots, d$ where d is the dimension of the input. The statement can be easily extended to methods that do not use attribution as well.

b) Input invariance: To illustrate using image classification problem, translating an image will also translate super-pixels demarcating the area that provides an explanation to the choice of classification correspondingly. Clearly, this property is desirable and has been proposed as an axiomatic invariance of a reliable saliency method. There has also been a study on the input invariance of some saliency methods with respect to translation of input $x \rightarrow x + c$ for some c [71]. Of the methods studied, gradients/sensitivity-based methods [128] and signal methods [72], [75] are input invariant while some attribution methods, such as integrated gradient [94], are not.

3) Interpretabilities by Utilities: The following utilities-based categorization of interpretability is proposed in [10].

a) Application-based: First, an evaluation is application-grounded if human A gives explanation X_A on a specific application, so-called the end-task (e.g., a doctor performs diagnosis) to human B, and B performs the same task. Then A has given B a useful explanation if B performs better in the task. Suppose A is now an ML model, then the model is highly interpretable if human B performs the same task with improved performance after given X_A. Some medical segmentation works will fall into this category as well, since the segmentation will constitute a visual explanation for further diagnosis/prognosis [144], [145] (also see other categories of the grand challenge). Such evaluation is performed, for example, in [43]. They proposed Grad-CAM applied on guided backpropagation (proposed in [75]) of AlexNet CNN and VGG. The produced visualizations are used to help human subjects in Amazon mechanical turks identify objects with higher accuracy in predicting VOC 2007 images. The human subjects achieved 61.23% accuracy, which is 16.79% higher than visualization provided by guided backpropagation.

b) Human-based: This evaluation involves real humans and simplified tasks. It can be used when, for some reasons or another, having human A give a good explanation X_A is challenging, possibly because the performance on the task cannot be evaluated easily or the explanation itself requires specialized knowledge. In this case, a simplified or partial problem may be posed and X_A is still demanded. Unlike the application-based approach, it is now necessary to look at X_A specifically for interpretability evaluation. Bigger pool of human subjects can then be hired to give a generic valuation to X_A or create a model answer $\hat{X}_A$ to compare X_A with, and then a generic valuation is computed.

Now, suppose A is an ML model, A is more interpretable compared to another ML model if it scores better in this generic valuation. In [146], an ML model is given a document containing the conversation of humans making a plan. The ML model produces a "report" containing relevant predicates (words) for the task of inferring what the final plan is. The metric used for interpretability evaluation is, for example, the percentage of the predicates that appear, compared to human-made report. We believe the format of human-based evaluation needs not be strictly like the above. For example, hybrid human and interactive ML classifiers require human users to nominate features for training [147]. Two different standard MLs can be compared to the hybrid, and one can be said to be more interpretable than another if it picks up features similar to the hybrid, assuming they perform at similarly acceptable level.

c) Functions-based: Third, an evaluation is functionally grounded if there exist proxies (which can be defined *a priori*) for evaluation, for example, sparsity [10]. Some articles [2], [5], [42]–[44], [96], [97], [144], and [145] use metrics that rely on this evaluation include many supervised learning models with clearly defined metrics such as: 1) dice coefficients (related to visual interpretability) and 2) attribution values, components of canonically transformed variables (see for example CCA) or values obtained from dimensionality reduction methods (such as components of principal components from PCA and their corresponding eigenvalues), where interpretability is related to the degree an object relates to a feature, for example, classification of a dog has high values in the feature space related to four limbs, shape of snout and paws, etc. Which suitable metrics to use are highly dependent on the tasks at hand.

III. XAI in Medical Field

ML has also gained traction recently in the medical field, with large volume of works on automated diagnosis, prognosis [150]. From the grand-challenge.org, we can see many different challenges in the medical field have emerged and galvanized researches that use ML and AI methods. Amongst

TABLE III
CATEGORIZATION BY THE ORGANS AFFECTED BY THE DISEASES. NEURO* REFERS TO ANY NEUROLOGICAL, NEURODEVELOPMENTAL, NEURODEGENERATIVE, ETC. DISEASES. THE ROWS ARE ARRANGED ACCORDING TO THE FOCUS OF THE INTERPRETABILITY AS THE FOLLOWING: APPL. = APPLICATION, METHOD. = METHODOLOGY, COMP. = COMPARISON

Appl.	brain, neuro* [48], [68], [131], [153] [132], [133], [136], [156]	breast [69], lung [6], [82], sleep [154], skin [155] others [106]
Method.	brain, neuro* [66], [67], [83], [91], [100] [114], [123], [135], [157]	breast [65], [70], [140], [141] skin [139], heart [124] others [44], [67], [138], [142]
Comp.	brain, neuro* [107], [158]	lung [93], sleep [159] skin [160], other [137]

successful DL models are [2], [5], using U-Net for medical segmentation. However, being a DL NN, U-Net is still a blackbox; it is not very interpretable. Other domain specific methods and special transformations (denoising etc.) have been published as well; consider for example [131] and many other works in MICCAI publications.

In the medical field, the question of interpretability is far from just intellectual curiosity. More specifically, it is pointed out that interpretabilities in the medical fields include factors other fields do not consider, including risk and responsibilities [21], [151], [152]. When medical responses are made, lives may be at stake. To leave such important decisions to machines that could not provide accountabilities would be akin to shirking the responsibilities altogether. Apart from ethical issues, this is a serious loophole that could turn catastrophic when exploited with malicious intent.

Many more works have thus been dedicated to exploring explainability in the medical fields [11], [20], [44]. They provide summaries of previous works [21] including subfield-specific reviews such as [25] for chest radiograph and sentiment analysis in medicine [161], or at least set aside a section to promote awareness for the importance of interpretability in the medical field [162]. In [163], it is stated directly that being a black box is a "strong limitation" for AI in dermatology, as it is not capable of performing customized assessment by certified dermatologist that can be used to explain clinical evidence. On the other hand, the exposition [164] argues that a certain degree of opaqueness is acceptable, that is, it might be more important that we produce empirically verified accurate results than focusing too much on how to the unravel the black-box. We recommend readers to consider them first, at least for an overview of interpretability in the medical field.

We apply categorization from the previous section to the ML and AI in the medical field. Table III shows categorization obtained by tagging: 1) how interpretability method is incorporated: either through direct application of existing methods, methodology improvements, or comparison between interpretability methods and 2) the organs targeted by the diseases for example, brain, skin, etc. As there is not yet a substantial number of significant medical researches that address interpretability, we will refrain from presenting any conclusive trend. However, from a quick overview, we see that the XAI research community might benefit from more studies comparing different existing methods, especially those with more informative conclusion on how they contribute to interpretability.

A. Perceptive Interpretability

Medical data could come in the form of traditional 2-D images or more complex formats such as NIFTI or DCOM which contain 3-D images with multiple modalities and even 4-D images which are time-evolving 3-D volumes. The difficulties in using ML for these data include the following. Medical images are sometimes far less available in quantity than common images. Obtaining these data requires consent from the patients and other administrative barriers. High-dimensional data also add complexity to data processing and the large memory space requirement might prevent data to be input without modification, random sampling or down-sizing, which may compromise analysis. Other possible difficulties with data collection and management include as left/right-censoring, patients' death due to unrelated causes or other complications etc.

When medical data is available, ground-truth images may not be "correct." Not only do these data require some specialized knowledge to understand, the lack of comprehensive understanding of biological components complicates the analysis. For example, ADC modality of MR images and the isotropic version of DWI are in some sense derivative, since both are computed from raw images collected by the scanner. Furthermore, many CT or MRI scans are presented with skull-stripping or other preprocessing. However, without a more complete knowledge of what fine details might have been accidentally removed, we cannot guarantee that an algorithm can capture the correct features.

1) Saliency: The following articles consist of direct applications of existing saliency methods. Chexpert [6] uses GradCAM for visualization of pleural effusion in a radiograph. CAM is also used for interpretability in brain tumor grading [153]. Tang *et al.* [68] uses guided Grad-CAM and feature occlusion, providing complementary heatmaps for the classification of Alzheimer's disease pathologies. Integrated gradient method and SmoothGrad are applied for the visualization of CNN ensemble that classifies estrogen receptor status using breast MRI [69]. LRP on DeepLight [48] was applied on fMRI data from Human Connectome Project to generate heatmap visualization. Saliency map has also been computed using primitive gradient of loss, providing interpretability to the NN used for electroencephalogram (EEG) sleep stage scoring [154]. There has even been a direct comparison between the feature maps within CNN and skin lesion images [155], overlaying the scaled feature maps on top of the images as a means to interpretability. Some images correspond to relevant features in the lesion, while others appear to explicitly capture artifacts that might lead to prediction bias.

The following articles are focused more on comparison between popular saliency methods, including their derivative/improved versions. Jansen *et al.* [159] trains an artificial NN for the classification of insomnia using physiological network (PN). The feature relevance scores are computed

from several methods, including DeepLIFT [57]. Comparison between four different visualizations is performed in [158]. It shows different attributions between different methods, and concluded that LRP and guided backpropagation provide the most coherent attribution maps in their Alzheimer's disease study. Basic tests on GradCAM and SHAP on dermoscopy images for melanoma classification are conducted, concluding with the need for significant improvements to heatmaps before practical deployment [160].

The following includes slightly different focus on methodological improvements on top of the visualization. Respond-CAM [44] is derived from [42] and [43], and provides a saliency map in the form of heat-map on 3-D images obtained from cellular electron cryo-tomography. High intensity in the heatmap marks the region where macromolecular complexes are present. Multilayer CAM (MLCAM) is introduced in [91] for glioma (a type of brain tumor) localization. Multiinstance (MI) aggregation method is used with CNN to classify breast tumor tissue microarray (TMA) image's for five different tasks [65], for example the classification of the histologic subtype. Super-pixel maps indicate the region in each TMA image where the tumor cells are; each label corresponds to a class of tumor. These maps are proposed as the means for visual interpretability. Also, see the activation maps in [66] where interpretability is studied by corrupting image and inspecting region of interest (ROI). The autofocus module from [67] promises improvements in visual interpretability for segmentation on pelvic CT scans and segmentation of tumor in brain MRI using CNN. It uses attention mechanism (proposed in [92]) and improves it with adaptive selection of scale with which the network "sees" an object within an image. With the correct scale adopted by the network while performing a single task, human observer analyzing the network can understand that a NN is properly identifying the object, rather than mistaking the combination of the object plus the surrounding as the object itself.

There is also a different formulation for the generation of saliency maps [70]. It defines a different softmax-like formula to extract signals from DNN for visual justification in classification of breast mass (malignant/benign). Textual justification is generated as well.

2) Verbal: In [82], a rule-based system could provide the statement "has asthma $\rightarrow$ lower risk," where risk here refers to death risk due to pneumonia. Likewise, Letham *et al.* [83] creates a model called Bayesian rule lists that provides such statements for stroke prediction. Textual justification is also provided in the LSTM-based breast mass classifier system [70]. The argumentation theory is implemented in the ML training process [156], extracting arguments or decision rules as the explanations for the prediction of stroke based on the asymptomatic carotid stenosis and risk of stroke (ACSRS) data set.

One should indeed look closer at the interpretability in [82]. Just as many MLs are able to extract some humanly nonintuitive pattern, the rule-based system seems to have captured the strange link between asthma and pneumonia. The link becomes clear once the actual explanation based on real situation is provided: a pneumonia patient which also suffers from asthma is often sent directly to the intensive care unit (ICU) rather than a standard ward. Obviously, if there is a variable ICU $=$ 0 or 1 that indicates admission to ICU, then a better model can provide more coherent explanation "asthma $\rightarrow$ ICU $\rightarrow$ lower risk." In the article, the model appears not to identify such variable. We can see that interpretability issues are not always clear-cut.

Several researches on VQA in the medical field have also been developed. The initiative by ImageCLEF [165], [166] appears to be at its center, though VQA itself has yet to gain more traction and successful practical demonstration in the medical sector before widespread adoption.

a) Challenges and future prospects: For perceptive interpretability in medical sector. In many cases, where saliency maps are provided, they are provided with insufficient evaluation with respect to their utilities within the medical practices. For example, when providing importance attribution to a CT scan used for lesion detection, are radiologists interested in heatmaps highlighting just the lesion? Are they more interested in looking for reasons why a hemorrhage is epidural or subdural when the lesion is not very clear to the naked eyes? There may be many such medically related subtleties that interpretable AI researchers may need to know about.

B. Interpretability via Mathematical Structure

1) Predefined Model: Models help with interpretability by providing a generic sense of what a variable does to the output variable in question, whether in medical fields or not. A parametric model is usually designed with at least an estimate of the working mechanism of the system, with simplification and based on empirically observed patterns. For example, Ulas *et al.* [131] uses kinetic model for the cerebral blood flow in $ml/100g/\text{min}$ with

$$\text{CBF} = f(\Delta M)\frac{6000\beta\,\Delta M \exp\left(\frac{\text{PLD}}{T_{1b}}\right)}{2\alpha T_{1b}(\text{SI}_{\text{PD}})\left(1-\exp\left(-\frac{\tau}{T_{1b}}\right)\right)} \tag{1}$$

which depends on perfusion-weighted image ΔM obtained from the signal difference between labeled image of arterial blood water treated with RF pulses and the control image. This function is incorporated in the loss function in the training pipeline of a fully CNN. At least, an interpretation can be made partially: the NN model is designed to denoise a perfusion-weighted image (and thus improve its quality) by considering CBF. How the network "understands" the CBF is again an interpretability problem of a NN which has yet to be resolved.

There is an inherent simplicity in the interpretability of models based on linearity, and thus they have been considered obviously interpretable as well; some examples include linear combination of clinical variables [100], metabolites signals for magnetic resonance spectroscopy (MRS), [106] etc. Linearity in different models used in the estimation of brain states is discussed in [107], including how it is misinterpreted. It compares what it refers to as forward and backward models and then suggested improvement on linear models. In [82], a logistic regression model picked up a relation

between asthma and lower risk of pneumonia death, that is, asthma has a negative weight as a risk predictor in the regression model. Generative discriminative machine (GDM) combines ordinary least square regression and ridge regression to handle confounding variables in Alzheimer's disease and schizophrenia data set [100]. GDM parameters are said to be interpretable, since they are linear combinations of the clinical variables. DL has been used for PET pharmacokinetic (PK) modeling to quantify tracer target density [132]. CNN has helped PK modeling as a part of a sequence of processes to reduce PET acquisition time, and the output is interpreted with respect to the golden standard PK model, which is the linearized version of simplified reference tissue model (SRTM). DL method is also used to perform parameters fitting for MRS [106]. The parametric part of the MRS signal model specified, $x(t) = \Sigma a_m x_m(t) e^{\Delta\alpha_m t + 2\pi i \Delta f_m t}$, consists of linear combination of metabolite signals $x_m(t)$. The article shows that the error measured in symmetric mean absolute percentage error (SMAPE) is smallest for most metabolites when their CNN model is used. In cases like this, clinicians may find the model interpretable as long as the parameters are well-fit, although the NN itself may still not be interpretable.

The models above use linearity for studies related to brain or neuro-related diseases. Beyond linear models, other brain and neuro-systems can be modeled with relevant subject-content knowledge for better interpretability as well. Segmentation task for the detection of brain midline shift is performed using using CNN with standard structural knowledge incorporated [133]. A template called model-derived age norm is derived from mean values of sleep EEG features of healthy subjects [157]. Interpretability is given as the deviation of the features of unhealthy subject from the age norm.

On a different note, RL has been applied to personalized healthcare. In particular, Zhu *et al.* [134] introduces group-driven RL in personalized healthcare, taking into considerations different groups, each having similar agents. As usual, Q-value is optimized with respect to policy π_θ, which can be qualitatively interpreted as the maximization of rewards over time over the choices of action selected by many participating agents in the system.

a) Challenges and future prospects: Models may be simplifying intractable system. As such, the full potential of ML, especially DNN with huge number of parameters, may be under-used. A possible research direction that taps onto the hype of predictive science is as the following: given a model, is it possible to augment the model with new, sophisticated components, such that parts of these components can be identified with (and thus interpreted as) new insights? Naturally, the augmented model needs to be comparable to previous models and shown with clear interpretation why the new components correspond to insighs previously missed. Do note that there are critiques against the hype around the potential of AI which we will leave to the readers.

2) Feature Extraction: Vanilla CNN is used in [142] but it is suggested that interpretability can be attained using a separable model. The separability is achieved by polynomial-transforming scalar variables and further processing, giving rise to weights useful for interpretation. In [123], fMRI is analyzed using correlation-based functional graphs. They are then clustered into super-graph, consisting of subnetworks that are defined to be interpretable. A convolutional layer is then used on the super-graph. For more references about NNs designed for graph-based problems, see the article's citations. The following are further subcategorization for methods that revolve around feature extraction and the evaluations or measurements (such as correlations) used to obtain the features, similar to the previous section.

a) Correlation: DWT-based method (discrete wavelet transform) is used to perform feature extraction before eventually feeding the EEG data (after a series of processings) into a NN for epilepsy classification [135]. A fuzzy relation analogous to correlation coefficient is then defined. Furthermore, as with other transform methods, the components (the wavelets) can be interpreted componentwise. As a simple illustration, the components for Fourier transform could be taken as how much certain frequency is contained in a time series. Zhang *et al.* [136] mentioned a host of wavelet-based feature extraction methods and introduced maximal overlap discrete wavelet package transform (MODWPT) also applied on EEG data for epilepsy classification.

Frame singular value decomposition (F-SVD) is introduced for classifications of electromyography (EMG) data [114]. It is a pipeline involving a number of processing that includes DWT, CCA, and SVD, achieving around 98% accuracies on classifications between amyotrophic lateral sclerosis, myopathy, and healthy subjects. Consider also CCA-based articles that are cited in the article, in particular citations 18–21 for EMG and EEG signals.

b) Clustering: VAE is used to obtain vectors in 64-D latent dimension to predict whether the subjects suffer from hypertrophic cardiomyopathy (HCM) [124]. A nonlinear transformation is used to create LE with two dimensions, which is suggested as the means for interpretability. Skin images are clustered [139] for melanoma classification using k-nearest-neighbor that is customized to include CNN and triplet loss. A queried image is then compared with training images ranked according to similarity measure visually displayed as query-result activation map pair.

t-SNE has been applied on human genetic data and shown to provide more robust dimensionality reduction compared to PCA and other methods [137]. Multiple maps t-SNE (mm-t-SNE) is introduced in [138], performing clustering on phenotype similarity data.

c) Sensitivity: Regression concept vectors (RCVs) is proposed along with a metric *Br* score as improvements to TCAV's concept separation [140]. The method is applied on breast cancer histopathology classification problem. Furthermore, unit ball surface sampling (UBS) metric is introduced [141] to address the shortcoming of *Br* score. It uses NNs for classification of nodules for mammographic images. Guidelinebased Additive eXplanation (GAX) is introduced in [93] for diagnosis using CT lung images. Its pipeline includes LIME-like perturbation analysis and SHAP. Comparisons are then made with LIME, Grad-CAM, and feature importance generated by SHAP.

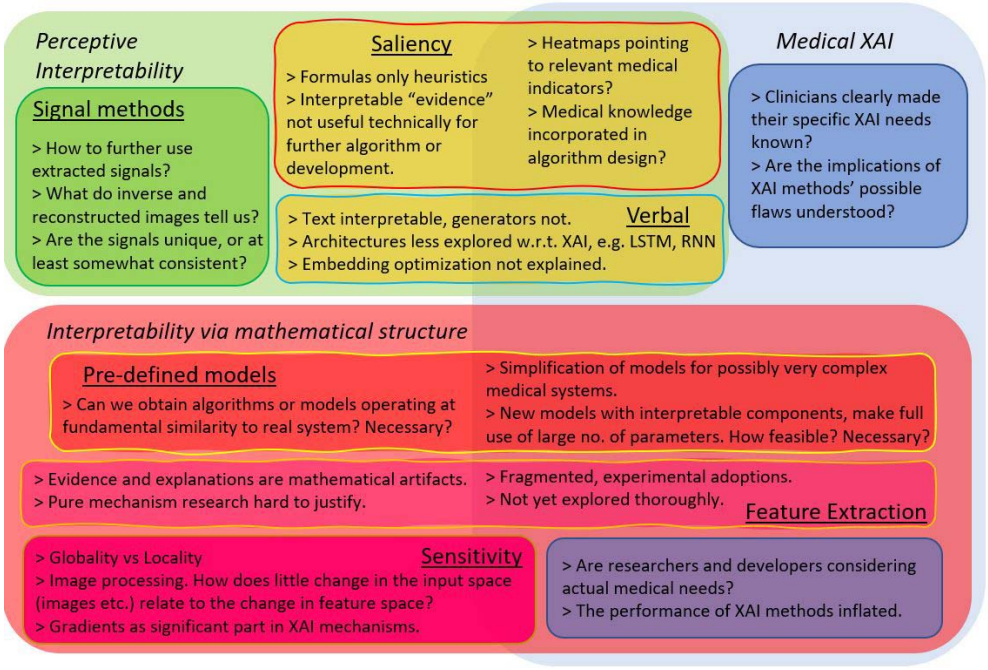

Fig. 6. Overview of challenges and future prospects arranged in a Venn diagram.

d) Challenges and future prospects: We observe popular uses of certain methods ingrained in specific sectors on the one hand and, on the other hand, emerging applications of sophisticated ML algorithms. As medical ML (in particular the application of recently successful DNN) is still a young field, we see fragmented and experimental uses of existing or customized interpretable methods. As medical ML research progresses, the tradeoff between many practical factors of ML methods (such as ease of use, ease of interpretation of mathematical structure possibly regarded as complex) and its contribution to the subject matter will become clearer. Future research and application may benefit from a practice of consciously and consistently extracting interpretable information for further processing, and the process should be systematically documented for good dissemination. Currently, with feature selections and extractions focused on improving accuracy and performance, we may still have vast unexplored opportunities in interpretability research.

C. Other Perspectives

1) Data-Driven: Case-based reasoning (CBR) performs medical evaluation (classifications etc.) by comparing a query case (new data) with similar existing data from a database. Lamy *et al.* [143] combines CBR with an algorithm that presents the similarity between these cases by visually providing proxies and measures for users to interpret. By observing these proxies, the user can decide to take the decision suggested by the algorithm or not. The article also asserts that medical experts appreciate such visual information with clear decision-support system.

D. Risk of Machine Interpretation in Medical Field

1) Jumping Conclusion: According to [82], logical statements such as has asthma → lower risk are considered interpretable. However, in the example, the statement indicates that a patient with asthma has lower risk of death from pneumonia, which might be strange without any clarification from the intermediate thought process. While human can infer that the lowered risk is due to the fact that pneumonia patients with asthma history tend to be given more aggressive treatment, we cannot always assume there is a similar humanly inferable reason behind each decision. Furthermore, interpretability method such as LRP, deconvolution, and guided backpropagation introduced earlier are shown to not work for simple model, such as linear model, bringing into question their reliability [60].

IV. Conclusion

We present a survey on interpretability and explainability of ML algorithms in general, and place different interpretations suggested by different research works into distinct categories. From general interpretabilities, we apply the categorization into the medical field. Some attempts are made to formalize interpretabilities mathematically, some provide visual explanations, while others might focus on the improvement in task performance after being given explanations produced by algorithms. At each section, we also discuss related challenges and future prospects. Fig. 6 provides a diagram that summarizes all the challenges and prospects.

A. Manipulation of Explanations

Given an image, a similar image can be generated that is perceptibly indistinguishable from the original, yet produces radically different output [95]. Naturally, its significance attribution and interpretable information become unreliable. Furthermore, explanation can even be manipulated arbitrarily [167]. For example, an explanation for the classification of a cat image (i.e., particular significant values that contribute

to the prediction of cat) can be implanted into the image of a dog, and the algorithm could be fooled into classifying the dog image as a cat image. The risk in medical field is clear: even without malicious, intentional manipulation, noises can render "explanations" wrong. Manipulation of algorithm that is designed to provide explanation is also explored in [168].

B. Incomplete Constraints

In [131], the loss function for the training of a fully convolutional network includes CBF as a constraint. However, many other constraints may play important roles in the mechanism of a living organ or tissue, not to mention applying kinetic model is itself a simplification. Giving an interpretation within limited constraints may place undue emphasis on the constraint itself. Other works that use predefined models might suffer similar problems [100], [106], [132].

C. Noisy Training Data

The so-called ground truths for medical tasks, provided by professionals, are not always absolutely correct. In fact, news regarding how AI beats human performance in medical imaging diagnosis [169] indicates that human judgment could be brittle. This is true even of trained medical personnel. This might give rise to the classic garbage-in-garbage-out situation.

The above risks are presented in large part as a reminder of the nature of automation. It is true that algorithms have been used to extract invisible patterns with some successes. However, one ought to view scientific problems with the correct order of priority. The society should not risk over-allocating resources into building machine and DL models, especially since due improvements to understanding the underlying science might be the key to solving the root problem. For example, higher quality MRI scans might reveal key information not "visible" with current technology, and many models built nowadays might not be very successful because there is simply not enough detailed information contained in currently available MRI scans.

D. Future Directions for Clinicians and Practitioners

Visual and textual explanation supplied by an algorithm might seem like the obvious choice; unfortunately, the details of decision-making by algorithms such as DNNs are still not clearly exposed. When an otherwise reliable DL model provides a strangely wrong visual or textual explanation, systematic methods to probe into the wrong explanations do not seem to exist, let alone methods to correct them. A specialized education combining medical expertise, applied mathematics, data science, etc. might be necessary to overcome this. For now, if "interpretable" algorithms are deployed in medical practices, human supervision is still necessary. Interpretability information should be considered nothing more than complementary support for the medical practices before there is a robust way to handle interpretability.

E. Future Directions for Algorithm Developers and Researchers

Before the blackbox is unblackboxed, machine decision always carries some exploitable risks. It is also clear that a unified notion of interpretability is elusive. For medical ML interpretability, more comparative studies between the performance of methods will be useful. The interpretability output such as heatmaps should be displayed and compared clearly, including poor results. In the best case scenario, clinicians and practitioners recognize the shortcomings of interpretable methods but have a general idea on how to handle them in ways that are suitable to medical practices. In the worst case scenario, the inconsistencies between these methods can be exposed. The very troubling trend of journal publications emphasizing good results is precarious, and we should thus continue interpretability research with a mindset open to evaluation from all related parties. Clinicians and practitioners need to be given the opportunity for fair judgment of utilities of the proposed interpretability methods, not just flooded with performance metrics possibly irrelevant to the adoption of medical technology.

Also, there may be a need to shift interpretability study away from algorithm-centric studies. An authoritative body setting up the standard of requirements for the deployment of model building might stifle the progress of the research itself, though it might be the most efficient way to reach an agreement. This might be necessary to prevent damages, seeing that even corporate companies and other bodies nonacademic in the traditional sense have joined the fray (consider health-tech start-ups and the implications). Acknowledging that machine and DL might not be fully mature for large-scale deployment, it might be wise to deploy the algorithms as a secondary support system for now and leave most decisions to the traditional methods. It might take a long time before humanity graduates from this stage, but it might be timely: we can collect more data to compare machine predictions with traditional predictions and sort out data ownership issues along the way.

ACKNOWLEDGMENT

The Alibaba-NTU Program is a collaboration between Alibaba and Nanyang Technological University, Singapore.

REFERENCES

[1] E.-J. Lee, Y.-H. Kim, N. Kim, and D.-W. Kang, "Deep into the brain: Artificial intelligence in stroke imaging," *J. Stroke*, vol. 19, no. 3, pp. 277–285, Sep. 2017.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, no. 3, pp. 234–241, Nov. 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[3] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Hum.-Comput. Stud.*, vol. 58, no. 6, pp. 697–718, Jun. 2003.

[4] L. Chen, P. Bentley, and D. Rueckert, "Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks," *NeuroImage, Clin.*, vol. 15, pp. 633–643, Jun. 2017.

[5] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," *CoRR*, vol. abs/1606.06650, pp. 6–7, Aug. 2016. [Online]. Available: http://arxiv.org/abs/1606.06650

[6] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *CoRR*, vol. abs/1901.07031, pp. 4–7, Jul. 2019. [Online]. Available: http://arxiv.org/abs/1901.07031

[7] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *CoRR*, vol. abs/1606.04797, pp. 7–9, Jun. 2016. [Online]. Available: http://arxiv.org/abs/1606.04797

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *CoRR*, vol. abs/1606.00915, pp. 7–11, Jun. 2016. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1606.html

[9] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, no. 1, p. 195, Dec. 2019.

[10] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*. [Online]. Available: http://arxiv.org/abs/1702.08608

[11] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," *CoRR*, vol. abs/1905.05134, pp. 1–12, May 2019. [Online]. Available: http://arxiv.org/abs/1905.05134

[12] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proc. ACM Conf. Comput. Supported Cooperat. Work (CSCW)*. New York, NY, USA: Association Computing Machinery, 2000, pp. 241–250.

[13] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Commun.*, vol. 10, no. 1, p. 1096, Dec. 2019.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association Computing Machinery, Aug. 2016, pp. 1135–1144.

[15] Z. C. Lipton, "The Mythos of model interpretability," *CoRR*, vol. abs/1606.03490, pp. 1–8, Mar. 2016. [Online]. Available: http://arxiv.org/abs/1606.03490

[16] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 210–215.

[17] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2018, pp. 80–89.

[18] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[19] S. R. Soekadar, N. Birbaumer, M. W. Slutzky, and L. G. Cohen, "Brain–machine interfaces in neurorehabilitation of stroke," *Neurobiol. Disease*, vol. 83, pp. 172–179, Nov. 2015.

[20] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 4, p. e1312, Jul. 2019.

[21] Y. Xie, G. Gao, and X. A. Chen, "Outlining the design space of explainable intelligent systems for medical diagnosis," *CoRR*, vol. abs/1902.06019, pp. 1–5, Mar. 2019. [Online]. Available: http://arxiv.org/abs/1902.06019

[22] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Comput. Appl.*, early access, Feb. 4, 2019, doi: 10.1007/s00521-019-04051-w.

[23] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Med.*, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[24] A. Fernandez, F. Herrera, O. Cordon, M. J. D. Jesus, and F. Marcelloni, "Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?" *IEEE Comput. Intell. Mag.*, vol. 14, no. 1, pp. 69–81, Feb. 2019.

[25] K. Kallianos *et al.*, "How far have we come? Artificial intelligence for chest radiograph interpretation," *Clin. Radiol.*, vol. 74, no. 5, pp. 338–345, May 2019.

[26] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[27] W. Samek, T. Wiegand, and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *CoRR*, vol. abs/1708.08296, pp. 1–6, Aug. 2017. [Online]. Available: http://arxiv.org/abs/1708.08296

[28] L. Rieger, P. Chormai, G. Montavon, L. K. Hansen, and K.-R. Müller, "Structuring neural networks for more explainable predictions," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham, Switzerland: Springer, 2018, pp. 115–131.

[29] S. Meacham, G. Isaac, D. Nauck, and B. Virginas, "Towards explainable AI: Design and development for explanation of machine learning predictions for a patient readmittance medical application," in *Intelligent Computing*, K. Arai, R. Bhatia, and S. Kapoor, Eds. Cham, Switzerland: Springer, 2019, pp. 939–955.

[30] J. Townsend, T. Chaton, and J. M. Monteiro, "Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3456–3470, Sep. 2020.

[31] (Oct. 2016). *Can We Open the Black Box of AI?* [Online]. Available: https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731

[32] B. Heinrichs and S. B. Eickhoff, "Your evidence? Machine learning algorithms for medical diagnosis and prediction," *Hum. Brain Mapping*, vol. 41, no. 6, pp. 1435–1444, Apr. 2020.

[33] M. Brundage *et al.*, "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," Eur. Commission, Brussels, Belgium, Tech. Rep., 2020. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai

[34] (Nov. 2019). *Ethics Guidelines for Trustworthy AI*. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[35] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable AI," in *Proc. CHI Conf. Hum. Factors Comput. Syst. (CHI)*. New York, NY, USA: Association Computing Machinery, 2019, pp. 1–15.

[36] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3319–3327.

[37] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2131–2145, Sep. 2019.

[38] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, p. e7, Nov. 2017.

[39] C. Olah *et al.*, "The building blocks of interpretability," Tech. Rep., Jan. 2020. [Online]. Available: https://distill.pub/2018/building-blocks/

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon *et al.*, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 4765–4774.

[41] A. Jacovi, O. S. Shalom, and Y. Goldberg, "Understanding convolutional neural networks for text classification," *CoRR*, vol. abs/1809.08037, pp. 3–7, Apr. 2018. [Online]. Available: http://arxiv.org/abs/1809.08037

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[43] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, pp. 1–21, Oct. 2016. [Online]. Available: http://arxiv.org/abs/1610.02391

[44] G. Zhao, B. Zhou, K. Wang, R. Jiang, and M. Xu, "Respond-CAM: Analyzing deep models for 3D imaging data by visualizations," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 485–492.

[45] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 2015.

[46] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.

[47] S. Becker, M. Ackermann, S. Lapuschkin, K. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *CoRR*, vol. abs/1807.03418, p. 3, Jul. 2018. [Online]. Available: http://arxiv.org/abs/1807.03418

[48] A. W. Thomas, H. R. Heekeren, K.-R. Müller, and W. Samek, "Analyzing neuroimaging data through recurrent deep learning models," *Frontiers Neurosci.*, vol. 13, p. 1321, Dec. 2019.

[49] L. Arras, F. Horn, G. Montavon, K. Müller, and W. Samek, "'What is relevant in a text document?': An interpretable machine learning approach," *CoRR*, vol. abs/1612.07843, pp. 1–17, Dec. 2016. [Online]. Available: http://arxiv.org/abs/1612.07843

[50] V. Srinivasan, S. Lapuschkin, C. Hellge, K.-R. Müller, and W. Samek, "Interpretable human action recognition in compressed domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1692–1696.
[51] O. Eberle, J. Buttner, F. Krautli, K.-R. Mueller, M. Valleriani, and G. Montavon, "Building and interpreting deep similarity models," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1, 2020.
[52] L. Hiley, A. Preece, Y. Hicks, S. Chakraborty, P. Gurram, and R. Tomsett, "Explaining motion relevance for activity recognition in video deep learning models," pp. 1–8, Mar. 2020, *arXiv:2003.14285*. [Online]. Available: https://arxiv.org/abs/2003.14285
[53] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, and K. Müller, "Interpreting the predictions of complex ML models by layer-wise relevance propagation," pp. 1–4, Nov. 2016, *arXiv:1611.08191*. [Online]. Available: https://arxiv.org/abs/1611.08191
[54] (2018). *Machine Learning and AI for the Sciences—Towards Interpretability*. [Online]. Available: http://www.heatmapping.org/slides/2018_WCCI.pdf
[55] (2016). *Deep Taylor Decomposition of Neural Networks*. [Online]. Available: http://iphome.hhi.de/samek/pdf/MonICML16.pdf
[56] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 9277–9286.
[57] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *CoRR*, vol. abs/1704.02685, pp. 1–8, Oct. 2017. [Online]. Available: http://arxiv.org/abs/1704.02685
[58] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *CoRR*, vol. abs/1702.04595, pp. 1–10, Feb. 2017. [Online]. Available: http://arxiv.org/abs/1702.04595
[59] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," *CoRR*, vol. abs/1804.00880, no. 2, pp. 1–8, Jun. 2018. [Online]. Available: http://arxiv.org/abs/1804.00880
[60] P. Kindermans *et al.*, "Learning how to explain neural networks: Patternnet and patternattribution," in *Proc. ICLR*, May 2018, pp. 1–11.
[61] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "Smoothgrad: Removing noise by adding noise," *CoRR*, vol. abs/1706.03825, pp. 1–8, Jun. 2017.
[62] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," in *Proc. 8th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.* Copenhagen, Denmark: Association Computational Linguistics, Sep. 2017, pp. 159–168.
[63] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," pp. 1–9, Jun. 2015, *arXiv:1506.02078*. [Online]. Available: https://arxiv.org/abs/1506.02078
[64] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. Robustness: Investigating medical imaging networks using adversarial examples," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 493–501.
[65] H. D. Couture, J. S. Marron, C. M. Perou, M. A. Troester, and M. Niethammer, "Multiple instance learning for heterogeneous images: Training a CNN for histopathology," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 254–262.
[66] X. Li, N. C. Dvornek, J. Zhuang, P. Ventola, and J. S. Duncan, "Brain biomarker interpretation in ASD using deep learning and fMRI," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 206–214.
[67] Y. Qin *et al.*, "Autofocus layer for semantic segmentation," *CoRR*, vol. abs/1805.08403, pp. 1–8, Jun. 2018. [Online]. Available: http://arxiv.org/abs/1805.08403
[68] Z. Tang *et al.*, "Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline," *Nature Commun.*, vol. 10, no. 1, p. 2173, Dec. 2019.
[69] Z. Papanastasopoulos *et al.*, "Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI," *Proc. SPIE*, vol. 11314, pp. 228–235, Mar. 2020.
[70] H. Lee, S. T. Kim, and Y. M. Ro, "Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, K. Suzuki *et al.*, Eds. Cham, Switzerland: Springer, 2019, pp. 21–29.
[71] P.-J. Kindermans *et al.*, *The (Un)reliability of Saliency Methods*. Cham, Switzerland: Springer, 2019, pp. 267–280.
[72] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR*, vol. abs/1311.2901, pp. 2–3, Nov. 2013. [Online]. Available: http://arxiv.org/abs/1311.2901
[73] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," *CoRR*, vol. abs/1412.0035, pp. 2–5, Nov. 2014. [Online]. Available: http://arxiv.org/abs/1412.0035
[74] A. Dosovitskiy and T. Brox, "Inverting convolutional networks with convolutional networks," *CoRR*, vol. abs/1506.02753, pp. 2–3, Apr. 2015. [Online]. Available: http://arxiv.org/abs/1506.02753
[75] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *CoRR*, vol. abs/1412.6806, pp. 1–10, Apr. 2015.
[76] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," Dept. d'Informatique Recherche Operationnelle, Univ. Montreal, Montreal, QC, Canada, Tech. Rep. 1341, Jun. 2009.
[77] A. M. Nguyen, J. Yosinski, and J. Clune, "Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks," *CoRR*, vol. abs/1602.03616, pp. 3–4, May 2016. [Online]. Available: http://arxiv.org/abs/1602.03616
[78] J. Yosinski, J. Clune, A. M. Nguyen, T. J. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *CoRR*, vol. abs/1506.06579, pp. 4–8, Jun. 2015. [Online]. Available: http://arxiv.org/abs/1506.06579
[79] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[80] R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen, "Ablation studies in artificial neural networks," *CoRR*, vol. abs/1901.08644, pp. 3–14, Feb. 2019. [Online]. Available: http://arxiv.org/abs/1901.08644
[81] R. Meyes, C. W. de Puiseau, A. Posada-Moreno, and T. Meisen, "Under the hood of neural networks: Characterizing learned representations by functional neuron populations and network ablations," pp. 1–8, May 2020, *arXiv:2004.01254*. [Online]. Available: https://arxiv.org/abs/2004.01254
[82] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: Association Computing Machinery, 2015, pp. 1721–1730.
[83] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *Ann. Appl. Statist.*, vol. 9, no. 3, pp. 1350–1371, Sep. 2015.
[84] I. Lage *et al.*, "An evaluation of the human-interpretability of explanation," *CoRR*, vol. abs/1902.00006, pp. 4–11, Jan. 2019. [Online]. Available: http://arxiv.org/abs/1902.00006
[85] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Faithful and customizable explanations of black box models," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.* New York, NY, USA: Association Computing Machinery, Jan. 2019, pp. 131–138.
[86] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Austin, TX, USA: Association Computational Linguistics, 2016, pp. 107–117.
[87] P. Guo, C. Anderson, K. Pearson, and R. Farrell, "Neural network interpretation via fine grained textual summarization," *CoRR*, vol. abs/1805.08969, pp. 2–5, Sep. 2018. [Online]. Available: http://arxiv.org/abs/1805.08969
[88] A. Agrawal *et al.*, "VQA: Visual question answering," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 4–31, May 2017.
[89] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2016, pp. 289–297.
[90] A. Das *et al.*, "Visual dialog," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 326–335.
[91] M. Izadyyazdanabadi *et al.*, "Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 300–308.

[92] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[93] P. Zhu and M. Ogino, "Guideline-based additive explanation for computer-aided diagnosis of lung nodules," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, K. Suzuki *et al.*, Eds. Cham, Switzerland: Springer, 2019, pp. 39–47.

[94] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[95] A. Ghorbani, A. Abid, and J. Y. Zou, "Interpretation of neural networks is fragile," in *Proc. AAAI*, Jul. 2019, pp. 1–7.

[96] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. ICML*, vol. 80, J. G. Dy and A. Krause, Eds. PMLR, 2018, pp. 2673–2682.

[97] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6078–6087.

[98] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–5.

[99] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *CoRR*, vol. abs/1703.00810, pp. 4–17, Apr. 2017. [Online]. Available: http://arxiv.org/abs/1703.00810

[100] E. Varol, A. Sotiras, K. Zeng, and C. Davatzikos, "Generative discriminative models for multivariate inference and statistical mapping in medical imaging," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 540–548.

[101] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," pp. 1–9, Nov. 2018, *arXiv:1610.01644*. [Online]. Available: https://arxiv.org/abs/1610.01644

[102] T. Hastie and R. Tibshirani, "Generalized additive models," *Stat. Sci.*, vol. 1, no. 3, pp. 297–310, Aug. 1986.

[103] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: Association for Computing Machinery, 2012, pp. 150–153.

[104] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: Association for Computing Machinery, 2013, p. 623.

[105] S. O. Arik and T. Pfister, "Attention-based prototypical learning towards interpretable, confident and robust deep neural networks," *CoRR*, vol. abs/1902.06292, pp. 3–6, Sep. 2019. [Online]. Available: http://arxiv.org/abs/1902.06292

[106] N. Hatami, M. Sdika, and H. Ratiney, "Magnetic resonance spectroscopy quantification using deep learning," *CoRR*, vol. abs/1806.07237, p. 3, Jun. 2018. [Online]. Available: http://arxiv.org/abs/1806.07237

[107] S. Haufe *et al.*, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, Feb. 2014.

[108] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nature Commun.*, vol. 8, no. 1, p. 13890, Apr. 2017.

[109] K. T. Schütt, M. Gastegger, A. Tkatchenko, and K.-R. Müller, *Quantum-Chemical Insights from Interpretable Atomistic Neural Networks*. Cham, Switzerland: Springer, 2019, pp. 311–330.

[110] C. Liaskos, A. Tsioliaridou, S. Nie, A. Pitsillides, S. Ioannidis, and I. F. Akyildiz, "An interpretable neural network for configuring programmable wireless environments," *CoRR*, vol. abs/1905.02495, pp. 2–3, May 2019. [Online]. Available: http://arxiv.org/abs/1905.02495

[111] B. Bede, "Fuzzy systems with sigmoid-based membership functions as interpretable neural networks," in *Fuzzy Techniques: Theory and Applications*, R. B. Kearfott, I. Batyrshin, M. Reformat, M. Ceberio, and V. Kreinovich, Eds. Cham, Switzerland: Springer, 2019, pp. 157–166.

[112] M. Kaiser, C. Otte, T. A. Runkler, and C. H. Ek, "Interpretable dynamics models for data-efficient reinforcement learning," *CoRR*, vol. abs/1907.04902, pp. 2–5, Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.04902

[113] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[114] A. Hazarika, M. Barthakur, L. Dutta, and M. Bhuyan, "F-SVD based algorithm for variability and stability measurement of bio-signals, feature extraction and fusion for pattern recognition," *Biomed. Signal Process. Control*, vol. 47, pp. 26–40, Jan. 2019.

[115] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2672–2680.

[116] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 214–223.

[117] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[118] Y. Zhu, S. Suri, P. Kulkarni, Y. Chen, J. Duan, and C.-C.-J. Kuo, "An interpretable generative model for handwritten digits synthesis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1910–1914.

[119] R. Krusinga, S. Shah, M. Zwicker, T. Goldstein, and D. W. Jacobs, "Understanding the (un)interpretability of natural image distributions using generative models," *CoRR*, vol. abs/1901.01499, pp. 5–9, Feb. 2019. [Online]. Available: http://arxiv.org/abs/1901.01499

[120] A. Karpathy. (2014). *t-SNE Visualization of CNN Codes*. [Online]. Available: https://cs.stanford.edu/people/karpathy/cnnembed

[121] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah. (2019). *Exploring Neural Networks With Activation Atlases*. [Online]. Available: https://distill.pub/2019/activation-atlas

[122] W. Ma, F. Cheng, Y. Xu, Q. Wen, and Y. Liu, "Probabilistic representation and inverse design of metamaterials based on a deep generative model with semi-supervised learning strategy," *Adv. Mater.*, vol. 31, no. 35, Aug. 2019, Art. no. 1901111.

[123] Y. Yan, J. Zhu, M. Duda, E. Solarz, C. Sripada, and D. Koutra, "Groupinn: Grouping-based interpretable neural network for classification of limited, noisy brain data," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 772–782.

[124] C. Biffi *et al.*, "Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling," *CoRR*, vol. abs/1807.06843, pp. 3–6, Jul. 2018. [Online]. Available: http://arxiv.org/abs/1807.06843

[125] M. Wang, D. Zhang, J. Huang, D. Shen, and M. Liu, "Low-rank representation for multi-center autism spectrum disorder identification," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 647–654.

[126] R. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," *CoRR*, vol. abs/1704.03296, pp. 1–8, Jan. 2017. [Online]. Available: http://arxiv.org/abs/1704.03296

[127] D. Alvarez-Melis and T. Jaakkola, "A causal framework for explaining the predictions of black-box sequence-to-sequence models," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Copenhagen, Denmark: Association Computational Linguistics, 2017, pp. 412–421.

[128] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, Aug. 2010.

[129] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1885–1894.

[130] C.-K. Yeh, J. S. Kim, I. E. Yen, and P. Ravikumar, "Representer point selection for explaining deep neural networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2018, pp. 9311–9321.

[131] C. Ulas, G. Tetteh, S. Kaczmarz, C. Preibisch, and B. H. Menze, "Deepasl: Kinetic model incorporated loss for denoising arterial spin labeled MRI via deep residual learning," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 30–38.

[132] C. J. Scott *et al.*, "Reduced acquisition time PET pharmacokinetic modelling using simultaneous ASL–MRI: Proof of concept," *J. Cerebral Blood Flow Metabolism*, vol. 39, no. 12, pp. 2419–2432, Dec. 2019.

[133] M. Pisov *et al.*, "Incorporating task-specific structural knowledge into CNNs for brain midline shift detection," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, K. Suzuki *et al.*, Eds. Cham, Switzerland: Springer, 2019, pp. 30–38.

[134] F. Zhu, J. Guo, Z. Xu, P. Liao, L. Yang, and J. Huang, "Group-driven reinforcement learning for personalized mhealth intervention," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 590–598.
[135] O. Kocadagli and R. Langari, "Classification of EEG signals for epileptic seizures using hybrid artificial neural networks based wavelet transforms and fuzzy relations," *Expert Syst. Appl.*, vol. 88, pp. 419–434, Dec. 2017.
[136] T. Zhang, W. Chen, and M. Li, "Classification of inter-ictal and ictal EEGs using multi-basis MODWPT, dimensionality reduction algorithms and LS-SVM: A comparative study," *Biomed. Signal Process. Control*, vol. 47, pp. 240–251, Jan. 2019.
[137] W. Li, J. E. Cerise, Y. Yang, and H. Han, "Application of t-SNE to human genetic data," *J. Bioinf. Comput. Biol.*, vol. 15, no. 4, Aug. 2017, Art. no. 1750017.
[138] W. Xu, X. Jiang, X. Hu, and G. Li, "Visualization of genetic disease-phenotype similarities by multiple maps t-SNE with Laplacian regularization," *BMC Med. Genomics*, vol. 7, no. 2, p. S1, 2014.
[139] N. C. F. Codella, C.-C. Lin, A. Halpern, M. Hind, R. Feris, and J. R. Smith, "Collaborative human-AI (CHAI): Evidence-based interpretable melanoma classification in dermoscopic images," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, D. Stoyanov *et al.*, Eds. Cham, Switzerland: Springer, 2018, pp. 97–105.
[140] M. Graziani, V. Andrearczyk, and H. Müller, "Regression concept vectors for bidirectional explanations in histopathology," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, D. Stoyanov *et al.*, Eds. Cham, Switzerland: Springer, 2018, pp. 124–132.
[141] H. Yeche, J. Harrison, and T. Berthier, "UBS: A dimension-agnostic metric for concept vector interpretability applied to radiomics," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, K. Suzuki *et al.*, Eds. Cham, Switzerland: Springer, 2019, pp. 12–20.
[142] A. E. U. Cerna *et al.*, "Interpretable neural networks for predicting mortality risk using multi-modal electronic health records," *CoRR*, vol. abs/1901.08125, pp. 3–6, Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.08125
[143] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artif. Intell. Med.*, vol. 94, pp. 42–53, Mar. 2019.
[144] Y. Choi, Y. Kwon, H. Lee, B. J. Kim, M. C. Paik, and J.-H. Won, "Ensemble of deep convolutional neural networks for prognosis of ischemic stroke," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, B. Menze, O. Maier, M. Reyes, S. Winzeck, and H. Handels, Eds. Cham, Switzerland: Springer, 2016, pp. 231–243.
[145] O. Maier and H. Handels, "Predicting stroke lesion and clinical outcome with random forests," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke Traumatic Brain Injuries*, A. Crimi, B. Menze, O. Maier, M. Reyes, S. Winzeck, and H. Handels, Eds. Cham, Switzerland: Springer, 2016, pp. 219–230.
[146] B. Kim, C. M. Chacha, and J. Shah, "Inferring robot task plans from human team meetings: A generative modeling approach with logic-based prior," in *Proc. 27th AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2013, pp. 1394–1400.
[147] J. Cheng and M. S. Bernstein, "Flock: Hybrid crowd-machine learning classifiers," in *Proc. 18th ACM Conf. Comput. Supported Cooperat. Work Social Comput.* New York, NY, USA: Association Computing Machinery, 2015, pp. 600–611.
[148] L. Kuhlmann *et al.*, "Epilepsyecosystem.Org: Crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG," *Brain*, vol. 141, pp. 2619–2630, Aug. 2018.
[149] M. Wiener, F. T. Sommer, Z. G. Ives, R. A. Poldrack, and B. Litt, "Enabling an open data ecosystem for the neurosciences," *Neuron*, vol. 92, no. 4, p. 929, Nov. 2016.
[150] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke Vascular Neurol.*, vol. 2, no. 4, pp. 230–243, Dec. 2017.
[151] C. K. Cassel and A. L. Jameton, "Dementia in the elderly: An analysis of medical responsibility," *Ann. Intern. Med.*, vol. 94, no. 6, pp. 802–807, Jun. 1981.
[152] P. Croskerry, K. Cosby, M. L. Graber, and H. Singh, *Diagnosis: Interpreting the Shadows*. London, U.K.: CRC Press, 2017.
[153] S. Pereira, R. Meier, V. Alves, M. Reyes, and C. A. Silva, "Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, D. Stoyanov *et al.*, Eds. Cham, Switzerland: Springer, 2018, pp. 106–114.
[154] A. Vilamala, K. H. Madsen, and L. K. Hansen, "Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring," in *Proc. IEEE 27th Int. Workshop Mach. Learn. for Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.
[155] P. Van Molle, M. De Strooper, T. Verbelen, B. Vankeirsbilck, P. Simoens, and B. Dhoedt, "Visualizing convolutional neural networks to improve decision support for skin lesion classification," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, D. Stoyanov *et al.*, Eds. Cham, Switzerland: Springer, 2018, pp. 115–123.
[156] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas, and C. Pattichis, "Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction," in *Proc. IEEE 19th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2019, pp. 817–821.
[157] H. Sun *et al.*, "Brain age from the electroencephalogram of sleep," *Neurobiol. Aging*, vol. 74, pp. 112–120, Feb. 2019.
[158] F. Eitel and K. Ritter, "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, K. Suzuki *et al.*, Eds. Cham, Switzerland: Springer, 2019, pp. 3–11.
[159] C. Jansen, T. Penzel, S. Hodel, S. Breuer, M. Spott, and D. Krefting, "Network physiology in insomnia patients: Assessment of relevant changes in network topology with interpretable machine learning models," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 29, no. 12, Dec. 2019, Art. no. 123129.
[160] K. Young, G. Booth, B. Simpson, R. Dutton, and S. Shrapnel, "Deep neural network or dermatologist?" in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, K. Suzuki *et al.*, Eds. Cham, Switzerland: Springer, 2019, pp. 48–55.
[161] C. Zucco, H. Liang, G. D. Fatta, and M. Cannataro, "Explainable sentiment analysis with applications in medicine," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 1740–1747.
[162] C. P. Langlotz *et al.*, "A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 NIH/RSNA/ACR/the academy workshop," *Radiology*, vol. 291, no. 3, pp. 781–791, Jun. 2019.
[163] A. Gomolin, E. Netchiporouk, R. Gniadecki, and I. V. Litvinov, "Artificial intelligence applications in dermatology: Where do we stand?" *Frontiers Med.*, vol. 7, p. 100, Mar. 2020.
[164] A. J. London, "Artificial intelligence and black-box medical decisions: Accuracy versus explainability," *Hastings Center Rep.*, vol. 49, no. 1, pp. 15–21, Jan. 2019.
[165] S. A. Hasan, Y. Ling, O. Farri, J. Liu, M. Lungren, and H. Müller, "Overview of the ImageCLEF 2018 medical domain visual question answering task," in *Proc. CLEF Working Notes, CEUR Workshop*, Avignon, France, Sep. 2018. [Online]. Available: http://ceur-ws.org
[166] A. Ben Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019," in *Proc. CLEF Working Notes, CEUR Workshop*, Lugano, Switzerland, Sep. 2019. [Online]. Available: http://ceurws.org
[167] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, "Explanations can be manipulated and geometry is to blame," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 13589–13600.
[168] H. Lakkaraju and O. Bastani, "'How do i fool you?' Manipulating user trust via misleading black box explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 79–85.
[169] Y. Liu *et al.*, "Detecting cancer metastases on gigapixel pathology images," *CoRR*, vol. abs/1703.02442, pp. 1–2, Mar. 2017. [Online]. Available: http://arxiv.org/abs/1703.02442